Synthetic Optimisation Techniques for Epidemic Disease Prediction Modelling

Terence Fusco, Yaxin Bi, Haiying Wang and Fiona Browne
Computer Science Research Institute, Ulster University, Shore Road, Newtownabbey, Antrim, Northern Ireland

Keywords: Optimisation, Over-sampling, Schistosomiasis, Synthetic Instance Generation, SMOTE, SMAC, SIMO.

Abstract: In this paper, research is presented for improving optimisation performance using sparse training data for disease vector classification. Optimisation techniques currently available such as Bayesian, Evolutionary and Global optimisation and are capable of providing highly efficient and accurate results however, performance potential can often be restricted when dealing with limited training resources. In this study, a novel approach is proposed to address this issue by introducing Sequential Model-based Algorithm Configuration(SMAC) optimisation in combination with Synthetic Minority Over-sampling Technique(SMOTE) for optimised synthetic prediction modelling. This approach generates additional synthetic instances from a limited training sample while concurrently seeking to improve best algorithm performance. As results show, the proposed Synthetic Instance Model Optimisation (SIMO) technique presents a viable, unified solution for finding optimum classifier performance when faced with sparse training resources. Using the SIMO approach, noticeable performance accuracy and f-measure improvements were achieved over standalone SMAC optimisation. Many results showed significant improvement when comparing collective training data with SIMO instance optimisation including individual performance accuracy increases of up to 46% and a mean overall increase for the entire 240 configurations of 13.96% over standard SMAC optimisation.

1 INTRODUCTION

Optimisation techniques are a sub-domain of machine learning that aim to discover optimal parameter and model-based conditions for a given dataset. The benefit of using an optimised approach is that it is possible to quickly eliminate inefficient and least effective algorithms from experiment conditions meaning more efficient testing of training data. Optimised solutions perform most effectively when using larger datasets due to speed capability for filtering vast amounts of data. These optimum processes involve using the most appropriate parameter variables or hyper-parameters in a training sample that enable the prediction model to best resolve the issue at hand. Hyper-parameter selection is an important tool in the optimisation process as it can concurrently target and improve weaknesses in a supplied data sample (Chan et al., 2013). The context of proposed optimisation approaches in this research are related to disease prediction modelling for future prevention and control purposes. The specific problem this work is focused on is the epidemic disease known as *schistosomiasis* and the host vector freshwater snail. *Schistosomiasis* disease is caused by parasitic worms and infection takes place when freshwater that has been contaminated by the snail comes into contact with humans, crops and cattle resulting in a detrimental effect on those infected. According to the World Health Organisation, in 2015 around 218 million people required preventative treatment and over 65 million people were treated for infection in the same year (W.H.O., 2016). The number of *schistosomiasis* infections has increased in recent years in many parts of Africa and Asia therefore, the need to provide early warning detection and prediction likelihood is imperative as a prevention and control tool. The aim of this research is to develop viable disease prediction models suitable for preventative measures to implemented by relevant health bodies. Once this information is provided to those communities at risk of disease outbreak, inhabitants can take known precautions to evade infection by avoiding unnecessary exposure to infested water bodies. The outcome of developing successful *schistosomiasis* prediction models for disease control purposes could drastically reduce the number of cases of people infected subsequently reducing costs associated with treatment and adverse effects on livestock and crops. The authen-

Fusco, T., Bi, Y., Wang, H. and Browne, F.
Synthetic Optimisation Techniques for Epidemic Disease Prediction Modelling.
DOI: 10.5220/0006823800950106
In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, pages 95-106
ISBN: 978-989-758-318-6

ticity and quantity of training data is of utmost importance when applying classification techniques for a particular disease research problem. A motivating factor for the proposed SIMO approach is the limited training resources available for application of algorithms to construct disease prediction models. A solution is required that modifies the existing training sample in a way that can improve classification potential without much distortion of the original data to avoid undermining prediction results. In the current state-of-art, automated optimisation techniques are providing significant performance improvements over standard applications leading to more technologically advanced approaches to a variety of research problems. Active learning approaches are becoming more prevalent in machine learning studies and are especially common with image classification research due to the nature of evolving discovery in that field. Optimisation links both automation and active learning methods to find those features and parameters which perform most favourably for a specific dataset (Settles, 2012). Successful optimisation application is often synonymous with larger datasets due to the ability to process information rapidly. Large reliable data sources can often be difficult to acquire in the epidemic disease exploration field, which has resulted in various techniques being constructed to amplify the sparse training data available. Current sampling methods applied for improving imbalanced training data address the issue using active learning and ensemble learning approaches which is an interesting method that achieved good results building on popular algorithms (Jian et al., 2016).

2 RELATED WORK

Optimisation techniques are becoming more prevalent for a variety of machine learning problems. Deep learning is one of the areas of interest which involves optimisation techniques such as neural network training and optimised machine learning algorithms (Bengio et al., 2015). Improvement of experiment efficiency and performance enhancement are principle factors in the application of these methods for use in the context of epidemic disease forecasting. This focus on optimisation research can prove to be a vital tool for rapid information sharing pertaining to a variety of disease monitoring studies. Constraints of this research regarding sparse training data prompted investigation of sampling methods that could improve class balance and increase machine learning potential (López et al., 2014). In addition to epidemiology studies, optimisation and parallel algorithm simulations have been previously applied to physics research with success and in particular the study of protein behaviours (Trebst et al., 2006). This *schistosomiasis* disease prediction research however, is more restrictive in terms of training data volume. Similar environment-based classification problems using sparse sample data can potentially benefit from findings in this work which face the same optimisation sample limitations. Opposing over-sampling and under-sampling techniques were considered and are assessed and expanded upon in this the following sections. Recent studies have compared real-world data and synthetic repository data for analysis of optimised active learning approaches which is a common method for optimised experiments (Krempl et al., 2015). The real-world data used in this research is used collectively and also as a base set for synthetic instance generation in order to assess the proposed optimised model in this study. Optimisation in many branches of machine learning requires an ever expanding number of training instances for comprehensive experiment conclusions and when this is not available Sequential model-based optimisation is an approach which applies algorithms in an iterative sequential order to achieve optimum learning conditions (Hutter et al., 2011). Another popular approach is Bayesian optimisation which employs an active learning procedure focusing on best performing algorithms during the optimisation process (Feurer et al., 2015).

2.1 Data Sampling

Data sampling or re-sampling of skewed data is a common technique used in machine learning and specifically when using real-world data. Class imbalance problems can frequently occur when using authentic environment data samples due to variations and density levels of spatial attributes. Over-sampling is a machine learning approach which uses additional sampling of instances in a supplied training set to increase the set size while balancing the data with increased minority classes. Under-sampling techniques are similar to over-sampling in that both methods share common aims but address the issue from different perspectives. Under-sampling is a contrapositive of over-sampling in that it reduces the size of a data sample by focusing on reduction of the majority class. A re-sampling approach was applied with this initial research to discover performance implications when using a limited training set. Similar conditions were applied to corresponding over-sampling experiments with the set used being the collective sample containing 223 instances and 8 attributes in total. The re-sampling method used provided a random sub-sample of the collective set using a bias to compensate for the minority class distribution. This was applied in fractions of the overall set for comparison purposes with under-samples of 100%, 80%, 60%, 40%, 20% and results recorded for analysis.

2.2 Synthetic Minority Over-sampling Technique

Synthetic Minority Over-sampling Technique (SMOTE) is a popular approach applied when using imbalanced sample data partly due to suitability for consecutive classification potential (Chawla et al., 2002). Increased training pools can provide improved classification potential therefore, over-sampling techniques were deemed to be the most appropriate sampling choice for this sparse data problem. SMOTE is a sampling approach aimed at increasing a dataset size with the purpose of improving minority class balance (Sáez et al., 2015). Synthetic instances are generated with minority bias as an alternative to over-sampling with replacement while also reducing the majority class hence increasing algorithm sensitivity to classifier assignment. For each instance x_i in the minority class, SMOTE searches the minority for the k nearest neighbours of x_i. One of these neighbours is selected as a seed sample $\hat{x}$. A random number between 0 and 1 denoted δ is chosen. The synthetic instance x_{new} is then created as shown in Equation 1.

$$x_{new} = x_i + (\hat{x} - x_i) \times \delta \tag{1}$$

2.3 Sequential Model-Based Algorithm Configuration

Sequential Model-Based Algorithm Configuration (SMAC) optimisation is a method that seeks to optimise model parameters to the ideal setting before classifier application. SMAC optimisation is similar in many ways to Bayesian optimisation in that it also uses a sequential approach with active learning for providing optimised algorithm conditions (Snoek et al., 2012). It aims to find the best performing model and parameter settings for a particular dataset in order to improve learning conditions for algorithms (Thornton et al., 2013). This is achieved using exploration of algorithm hyper-parameter space and includes examining new algorithms for performance analysis.

2.4 Research Issues

An issue that often arises when using automated optimisation processes concerns performance potential when using a sparse dataset for learning. Limited training resources can reduce effectiveness of optimisation capability and restrict potential for some automated techniques to be considered. The proposed model in this paper focuses on development of environment-based prediction models and issues surrounding the perceived lack of real-world data for modelling of vector-borne disease risk. In the data samples used in this work, there is evidence of class imbalance, which can be detrimental to the classification and prediction process (He et al., 2008). Real-world data composition in general terms tends to contain unequally represented class categories. Training samples used encompass a six-year period from 2003-2009 around the Dongting Lake area in Hunan Province, China. From initial analysis and pre-experiment study phases, vector classes were identified that were unequally represented. This imbalance can be due to a number of environment variables at the time of collection and is common issue with many real-world environment samples. Over-fitting can occur when classifying imbalanced data due to a predominant class in the set rendering classifier tendency to assign that class label to new instances. A simple solution for addressing class imbalance is to acquire additional data, which would increase the training pool and vary the class distribution. This is the most apparent approach when using a sparse data sample although difficulty lies in acquiring field survey information with corresponding freshwater snails. Lack of data and difficulty in accessing new samples is a significant component of the research problem being addressed with this work.

Development of viable disease prediction models can be restricted by a lack of field-survey data samples particularly in the case of vector-borne disease. Substantial collections of earth observation data have become more accessible in recent times however, corresponding vector distribution data has proven challenging to collect on a large scale therefore, optimisation approaches were investigated to maximise potential of classifier performance using limited data resources (Corne and Reynolds, 2010). Motivation for this work focuses on providing early warning information to at-risk communities that can help with prevention and control of *schistosomiasis* outbreak and the destructive effects of transmission in local communities. Successful results and improved optimisation performance of proposed SIMO method will inform future research on optimal synthetic instance

Table 1: Raw Data Snapshot.

AREA	SD	TCB	TCG	TCW	MNDWI	NDMI	NDVI	NDWI
N49	0.03	0.26	0.13	-0.10	-0.58	0.03	0.60	-0.61
N60	0.02	0.29	0.10	-0.09	-0.45	-0.01	0.43	-0.45
N74	0.10	0.52	0.08	-0.07	-0.24	0.00	0.18	-0.24
N75	2.26	0.21	0.07	-0.03	-0.27	0.11	0.32	-0.37
N76	0.37	0.41	0.11	-0.15	-0.47	-0.09	0.32	-0.40
N77	0.08	0.21	0.13	-0.01	-0.30	0.29	0.56	-0.54

dimensions and model parameters for classification and prediction modelling. This can contribute to the advancement in optimisation and over-sampling approaches in any further experiment capacity in this epidemiology domain.

Modified forms of SMOTE over-sampling and SMAC optimisation currently exist and are useful tools for many research problems however, the proposed SIMO method provides a unified approach combining the two techniques in order to find optimum classifier performance which includes the optimum performing synthetic sample increase quantity. The research presented does not provide incontrovertible evidence of impending disease outbreak but rather the most informed advice for monitoring and control to present to health agencies dealing with public health risks.

3 EXPERIMENT MATERIALS

Experiment data used in this paper is supplied by research partners at the European Space Agency (ESA) in conjunction with the Academy of Opto-electronics in Beijing, China. ESA partners provided satellite images over requested spatio-temporal parameters which was then used for environment feature extraction by research partners at the Chinese Academy of Sciences(CAS). Feature extraction was conducted using spectral and spatial software for high resolution image processing which provided raw labelled environment values. This data was then presented and processed before being deemed experiment ready. The study area on which all experiments are based is the Dongting Lake area of Hunan Province, China as shown in Figure 1.

3.1 Training Data

Training data was provided using a combination of satellite information and environment feature extraction techniques which was then presented in a raw data format before preprocessing for experiment purposes. A snapshot of the training sample is provided in Table 1. Training data used in these experiments is a collective sample ranging from 2003-2009 containing 223 instances with eight attributes. The environment attributes used in all experiments are as follows:

- **TCB** - Tasselled Cap Brightness (soil)
- **TCG** - Tasselled Cap Greenness (vegetation)
- **TCW** - Tasselled Cap Wetness (soil and moisture)
- **MNDWI** - Modified Normalised Difference Water Index (Water Index)
- **NDMI** - Normalised Difference Moisture Index (soil moisture)
- **NDVI** - Normalised Difference Vegetation Index (green vegetation)
- **NDWI** - Normalised Difference Water Index (water index)

The theory and rationale reinforcing proposed prediction models is that using satellite data and corresponding field-survey samples can help with making informed prediction models for application with future satellite extracted environment information used for training successful prediction models. The proposed synthetic optimsation method in this paper can assist in this research aim by assessing optimal classification parameters while evaluating synthetic instance generation viability on a sparse sample. The triumvirate of experts involved in this three-pronged research project are briefly described in Figure 2.

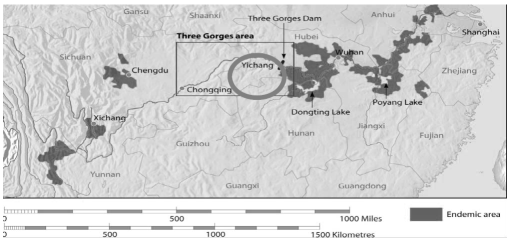


Figure 1: Study Area Map.

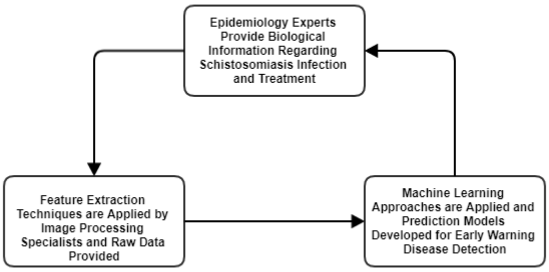

Epidemiology Experts Provide Biological Information Regarding Schistosomiasis Infection and Treatment
Feature Extraction Techniques are Applied by Image Processing Specialists and Raw Data Provided
Machine Learning Approaches are Applied and Prediction Models Developed for Early Warning Disease Detection

Figure 2: Research Partnership Components.

4 METHODOLOGY

To evaluate under-sampling of the sparse training set, a selection of established algorithms were applied having performed well in many fields of epidemic disease detection research to date (Lin et al., 2011). These included Naive Bayes, J48, SVM and MLP and results are assessed using classification accuracy with corresponding sample size as presented in Figure 4. Initial optimisation experiments were conducted on the collective training data to give the greatest data pool from the sparse samples for optimisation to take place. SMAC optimisation was applied with the top ten performing configurations being displayed for use in the next stage of testing (Kotthoff et al., 2016). Table 2 shows experiment duration results ranging from 1-24 hours of SMAC optimisation application with the collective training pool of 223 instances over a number of years from the Dongting Lake area of Hunan Province, China. At the end of each selected time period, the optimum performance algorithm together with weighted f-measure and classification accuracy findings were recorded. Each duration interval provided best performing algorithm results in terms of weighted f-measure and classification accuracy to provide a comprehensive algorithm analysis rather than classification accuracy metrics alone.

4.1 Synthetic Instance Model Optimisation

The proposed approach of this research is to implement Sequential Model-based Algorithm Configuration(SMAC) while simultaneously introducing an amplified number of synthetic instances using Synthetic Minority Over-Sampling Technique(SMOTE) to improve training potential with optimisation performance. In implementing this proposed Synthetic Instance Model Optimisation(SIMO), the aim is to increase performance of the optimised algorithm used to achieve greatest results. The success of this proposed method could alleviate the need to conduct much of the expensive and time-intensive field survey research that is required in order to make confident classification and prediction of the disease vector density and distribution. Results of this research will enable discovery of those classifiers which perform better with larger training sets of data and identify those poorly performing classifiers whose performance diminishes when increased synthetic instances are added. This information can be utilised for applying optimisation methods with future predictions.

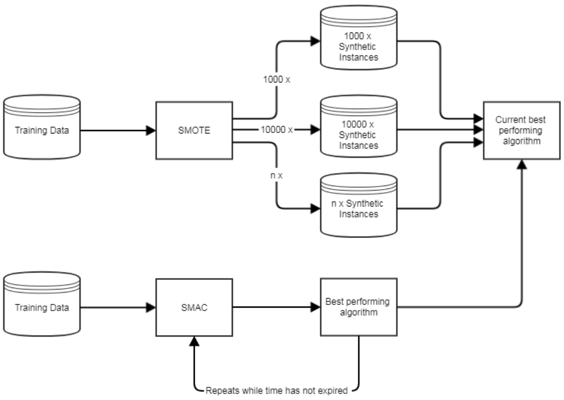

Training Data
SMOTE
1000 x
10000 x
n x
1000 x Synthetic Instances
10000 x Synthetic Instances
n x Synthetic Instances
Current best performing algorithm
Training Data
SMAC
Best performing algorithm
Repeats while time has not expired

Figure 3: SIMO Process Diagram.

Parallel algorithm configuration processing is a concept associated with optimisation and has been applied with success in the bioinformatics domain (Hutter et al., 2012). It is the assertion of this study that using optimised model processes in parallel with contributory sample balance improvement methods can significantly improve optimum performance potential of a sparse sample. The proposed SIMO model was constructed using a combined process involving SMOTE over-sampling and SMAC optimisation approaches. Combining approaches when optimising provides scope for improvement and can utilise the positive aspects of each respective technique. This SIMO approach is in essence an active learning approach which implements optimisation operations to both simulated instance sampling volume and model configuration selection.

The following stages describe the Synthetic Instance Model Optimisation (SIMO) process that are followed:

- SMAC optimisation is tested manually with the collective real-world training sample with duration intervals ranging from 1-24 hours.
- SMAC optimisation is then applied in conjunction with generated SMOTE synthetic data simulation.
- The top ten best performing algorithms from each duration interval are then applied sequentially with synthetically generated instances for performance analysis.

Table 2: Benchmark Optimisation Results.

NumHrs	Algorithm	WeightedF	Acc%
1	RandomTree	0.982	98.2
2	J48	0.663	69.1
3	Logistic	0.583	66.4
4	OneR	0.746	77.6
5	RandomTree	0.982	98.2
6	RandomTree	0.982	98.2
7	OneR	0.991	99.1
8	Logistic	0.991	99.1
9	Bagging	0.622	68.2
10	Adaboost	0.605	59.2
11	Vote	0.609	65.5
12	RandomTree	0.559	67.7
13	Logistic	0.59	66.8
14	OneR	0.62	69.1
15	Bagging	0.722	74.4
16	Logistic	0.66	71.3
17	RandomSubSpace	0.622	62.8
18	RandomSubSpace	0.599	62.8
19	RandomSubSpace	0.599	60.1
20	LWL	0.702	73.1
21	LWL	0.721	74.9
22	OneR	0.684	71.7
23	LMT	0.555	67.7
24	OneR	0.684	71.7

- Both approaches are then unified into a single optimisation process with the objective of providing optimised synthetic instance generation models.

A model diagram of proposed SIMO approach is shown in Figure 3 and shows the concurrent process with training data being introduced to both SMOTE and SMAC techniques before beginning the unified SIMO approach. The experiment process involves running SMAC optimisation for every hour ranging from 1-24 hours to assess performance of optimised techniques when applied with authentic sample data. Results of these initial tests were recorded and analysed for research purposes. Subsequently the SIMO unified approach was applied with synthetically generated data based on the original sample ranging from 1000, 5000 and 10,000 instance gamut. For each of these synthetic sets, the first ten recorded optimised results were extracted and contrasted with the performance from the original data classification to assess effectiveness. During each of the experiment phases, a SMOTE Equilibrium approach was applied with increasing sample magnitude to appropriately assess the effects of the proposed synthetic data simulation approach.

5 RESULTS

From initial results applied with an under-sampling technique in Figure 4, a gradual decline in performance accuracy it is noticeable with increases from 20% to 100% of the full sample size when under-sampling bias is implemented. This was expected from under-sampling of an already limited data pool but nonetheless contributed information of interest for assessing classifier behaviour with each batch increase. The performance of J48 decision tree decreased most significantly in terms of accuracy while MLP provided a performance gain between 20% and 100% sample size which can be factored into any future experiment thought process.

In relation to graphical representations in Figure 5, a selection of results are presented to show optimisation performance from novel SIMO method in comparison with collective training optimisation configurations. In each of the hourly configuration accuracy results, the original collective data sample is denoted using *C* with synthetic instance volume represented by *S* followed by instance number in 1000, 5000 and 10,000 gamuts. Figure 5 shows classification accuracy improvements in the majority of cases with increased synthetic instance simulation signifying optimisation performance improvements which shows scalability potential of the sparse training set. Similarly in Figures 6, significant f-measure performance improvements are noticeable when increasing synthetic instances to the optimisation process across the vast majority of models. These results using classification accuracy and f-measure metrics over a number of optimisation time intervals, help to reinforce the necessity of proposed SIMO model as an effective tool for improving epidemic risk prediction modelling when using sparse sample data.

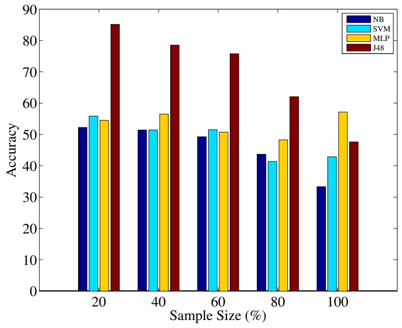
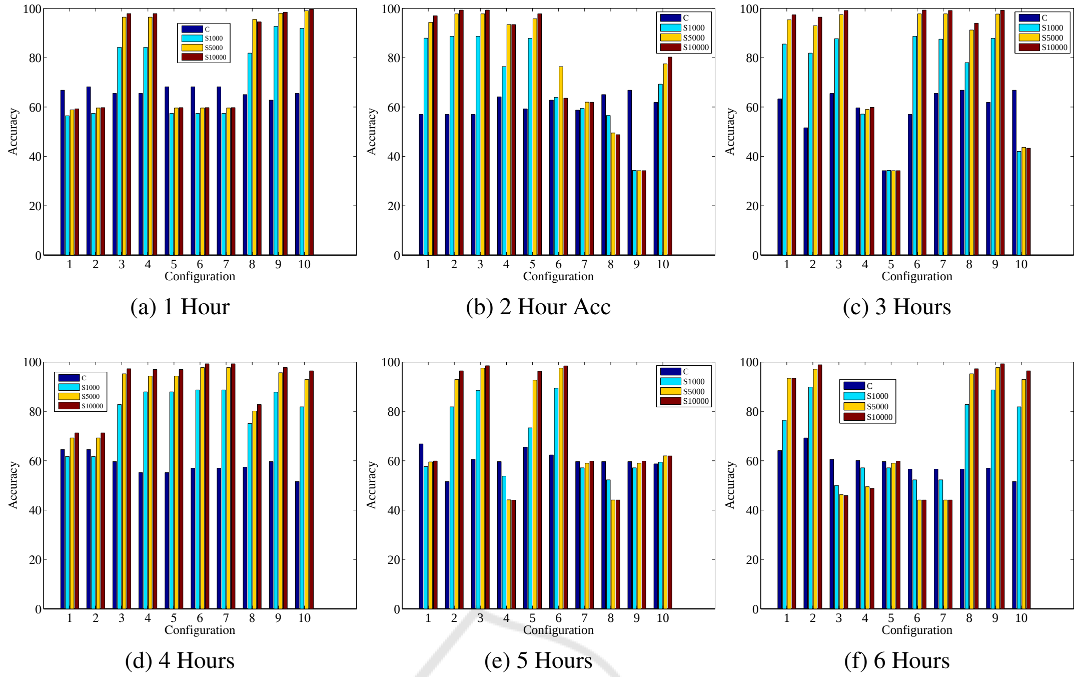


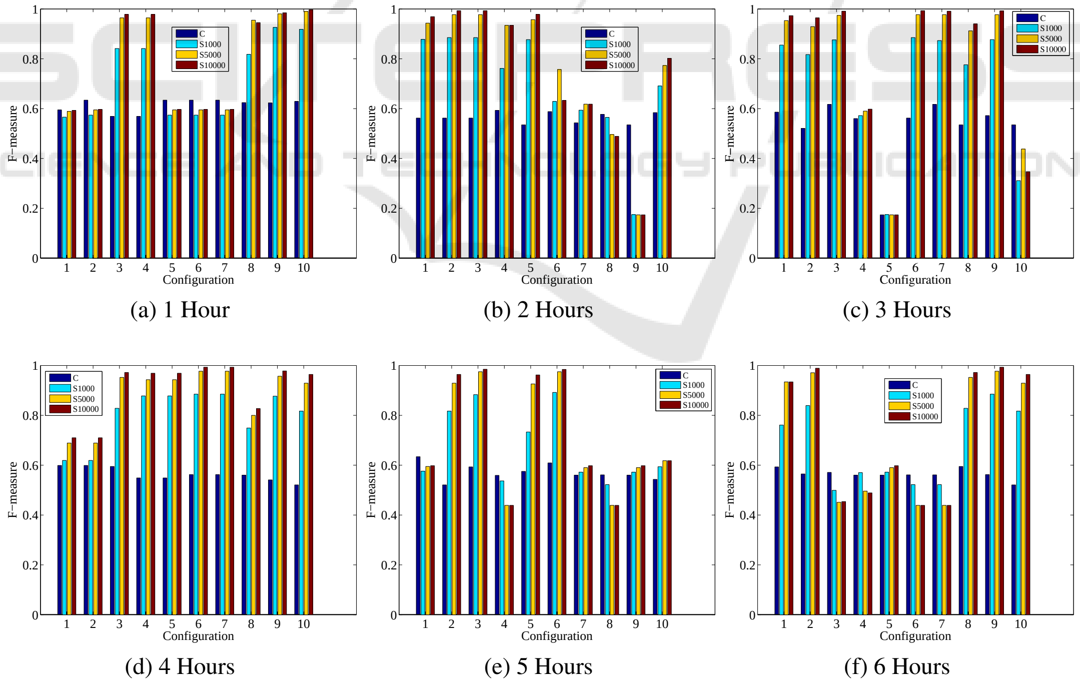
Figure 4: Under-sampling Results.



(a) 1 Hour (b) 2 Hour Acc (c) 3 Hours

(d) 4 Hours (e) 5 Hours (f) 6 Hours

Figure 5: SIMO Accuracy 1-6.



(a) 1 Hour (b) 2 Hours (c) 3 Hours

(d) 4 Hours (e) 5 Hours (f) 6 Hours

Figure 6: SIMO FMeasure Results 1-6.

5.1 Discussion

The method proposed in this paper is constructed using optimisation techniques in tandem with synthetic instance generation methods. The aim of this work is to find optimal conditions both in terms of parameter settings and instance simulation volume for making accurate classification of SD as well as discovery of environment attribute influence on SD levels. The predicted hypotheses of this work was that using the proposed SIMO method could improve accuracy and f-measure performance with synthetic instance optimisation over standalone SMAC optimisation during empirical experiments ranging in a 24-hour temporal parameter setting. The expectation from results using proposed SIMO approach is that performance improvement should be evident with each synthetic instance increment in comparison with optimised collective sample performance. This should be replicated with both accuracy and f-measure metrics in the main with some potential individual exceptions that will be identified for further analysis as is the case in Table 2 for hour 3. In Figures 5 and 6, some of these results are presented affirming the initial prediction both in terms of accuracy and f-measure metrics. This validation has rendered the SIMO model an effective performance enhancing model suitable for use when applying optimisation approaches to a sparse training sample. There are however some anomalous results with a number of poor optimisation performances observed when applied over a longer period of time compared with shorter experiment durations. These results require further investigation as to why performance was so poor with certain parameter settings and what the optimum classifiers from the poorest performing years were for future considerations.

6 CONCLUSIONS

In this study a novel SIMO method was presented using a hybrid approach incorporating SMAC optimsation and SMOTE instance generation with the aim of evaluating and assessing optimal instance generation volume and parameter settings for optimised classification. In summary, current findings have identified optimal parameter settings and classifiers for a range of duration intervals providing a knowledge base for future optimisation experiments in this field. Individual classifier performance can now be correctly distinguished as that which performs best with reduced or increased optimisation time periods. This information is indicative of each algorithm's potential for suitability with more machine intensive problems such as deep structured learning studies and can eliminate certain algorithms from future SIMO prediction training. In each of the examples in Figure 5 and 6, there is evidence of increasing accuracy and f-measure performance in the majority of cases which is positive for insight when building future predictive models. Another example is shown in Figure 10(a) with 13 hours of optimisation providing an average increase of **27.2%** on standard SMAC implementation with 9 out of 24 configurations having more than **15%** average accuracy increase. In terms of f-measure, 25% of average configuration increases resulted in more than **2.5** f-measure improvement with total average increase across all results of **0.18** and a high average increase of **3.4** when optimising for 4 hours with **0.43** increases in some cases. In Table 2, results show high frequency of certain algorithms such as OneR providing optimum performance in 5 of the 24-hour intervals with other similar regularity from Random Tree and Logistic Regression providing most accurate performance levels with accuracy in the high **90%** range. These classifiers indicate optimal suitability for use with this research problem and provide a basis for future baseline experiments with the novel SIMO model. The analysis factors that require further assessment based on all results will contrast the exploration and exploitation benefits that is, determining which performance level provides the greatest improvement while remaining efficient and maintaining data authenticity. The next phase of validating this method will involve empirical evaluation of alternative sampling methods and representative datasets for comparative performance analysis.

ACKNOWLEDGEMENTS

Training data is supplied by research partners at The European Space Agency in conjunction with The Academy of Opto-electronics, CAS, China.

REFERENCES

Bengio, Y., Goodfellow, I. J., and Courville, A. (2015). Optimization for training deep models. *Deep Learning*, pages 238–290.

Chan, S., Treleaven, P., and Capra, L. (2013). Continuous hyperparameter optimization for large-scale recommender systems. In *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, pages 350–358.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, pages 321–357.

Corne, D. W. and Reynolds, A. P. (2010). Optimisation and generalisation: Footprints in instance space. In *Lecture Notes in Computer Science*, volume 6238 LNCS, pages 22–31.

Feurer, M., Springenberg, J. T., and Hutter, F. (2015). Initializing Bayesian Hyperparameter Optimization via Meta-Learning. *Proceedings of the 29th Conference on Artificial Intelligence (AAAI 2015)*, pages 1128–1135.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, pages 1322–1328.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *Lecture Notes in Computer Science*, volume 6683 LNCS, pages 507–523.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2012). Parallel algorithm configuration. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7219 LNCS, pages 55–70.

Jian, C., Gao, J., and Ao, Y. (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, 193:115–122.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K. (2016). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 17:1–5.

Krempl, G., Kottke, D., and Lemaire, V. (2015). Optimised probabilistic active learning (OPAL): For fast, non-myopic, cost-sensitive active classification. *Machine Learning*, 100(2-3):449–476.

Lin, Y. L., Hsieh, J. G., Wu, H. K., and Jeng, J. H. (2011). Three-parameter sequential minimal optimization for support vector machines. *Neurocomputing*, 74(17):3467–3475.

López, V., Triguero, I., Carmona, C. J., García, S., and Herrera, F. (2014). Addressing imbalanced classification with instance generation techniques: IPADE-ID. *Neurocomputing*, 126:15–28.

Sáez, J. A., Luengo, J., Stefanowski, J., and Herrera, F. (2015). SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291(C):184–203.

Settles, B. (2012). Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

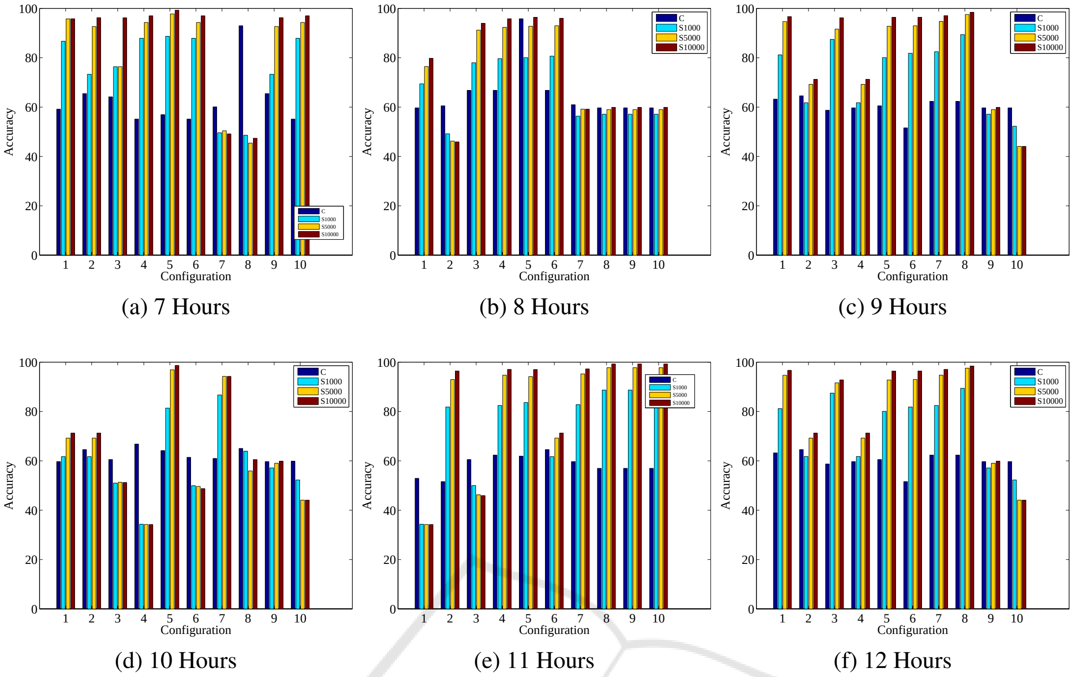
Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Adv. Neural Inf. Process. Syst. 25*, pages 1–9.

Thornton, C., Hutter, F., Hoos, H. H., Leyton-Brown, K., and Chris Thornton, Frank Hutter, Holger H. Hoos, K. L.-B. (2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 847–855.

Trebst, S., Troyer, M., and Hansmann, U. H. E. (2006). Optimized parallel tempering simulations of proteins. *Journal of Chemical Physics*, 124(17).

W.H.O. (2016). Schistosomiasis. Website. last checked: 21.04.2018.

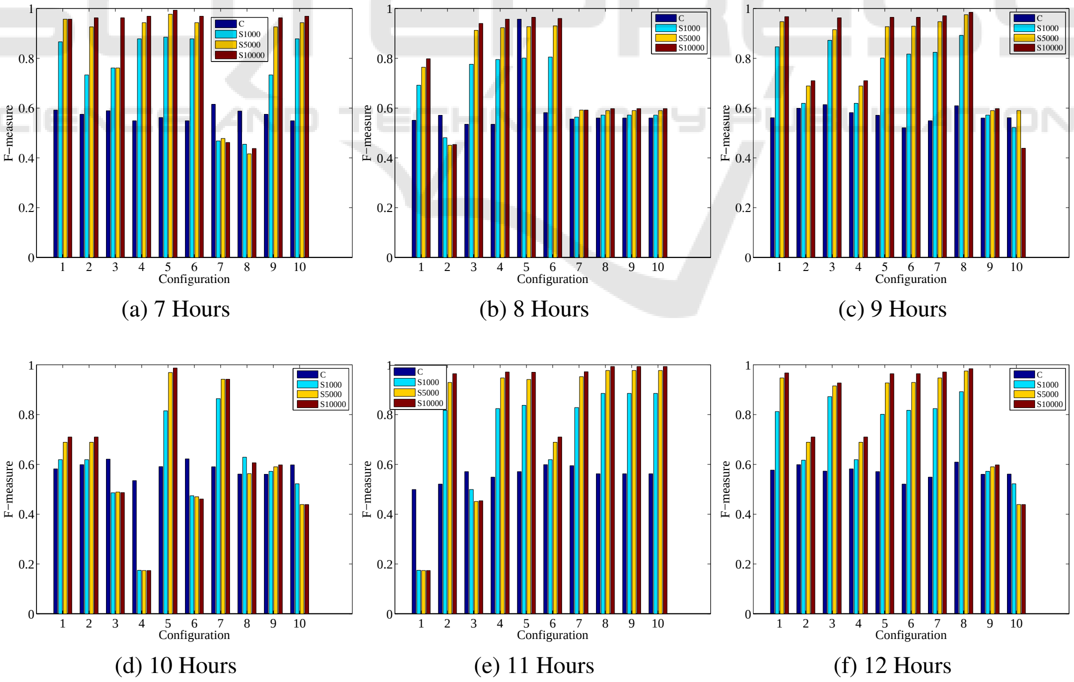
APPENDIX



(a) 7 Hours (b) 8 Hours (c) 9 Hours

(d) 10 Hours (e) 11 Hours (f) 12 Hours

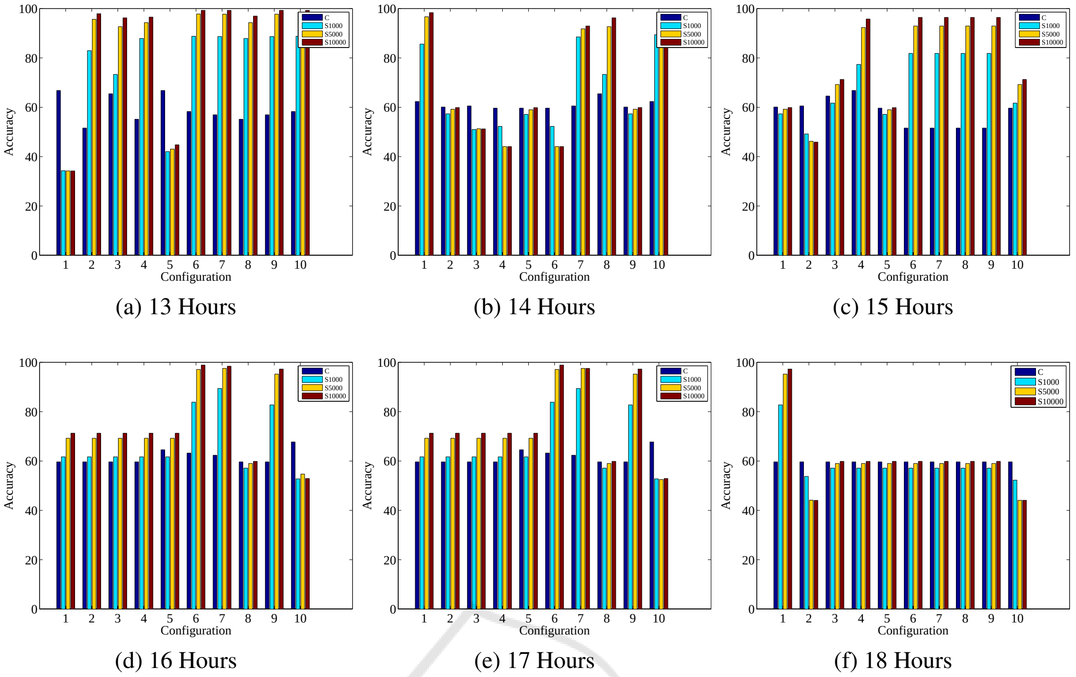
Figure 7: SIMO Accuracy Results 7-12.



(a) 7 Hours (b) 8 Hours (c) 9 Hours

(d) 10 Hours (e) 11 Hours (f) 12 Hours

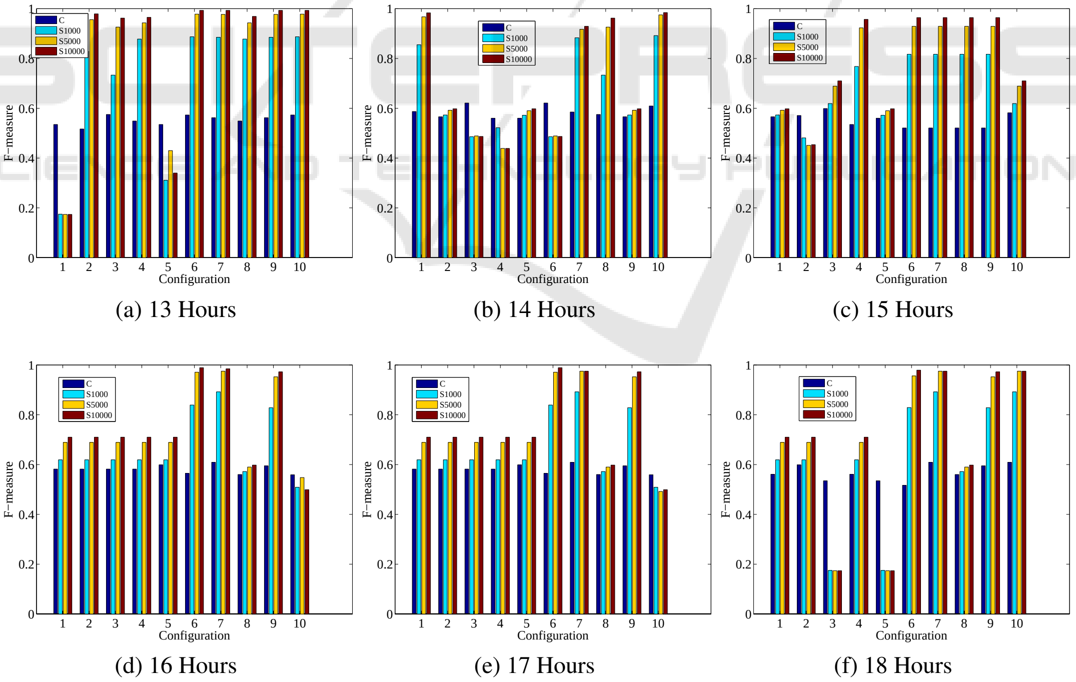
Figure 8: SIMO FMeasure Results 7-12.



(a) 13 Hours (b) 14 Hours (c) 15 Hours

(d) 16 Hours (e) 17 Hours (f) 18 Hours

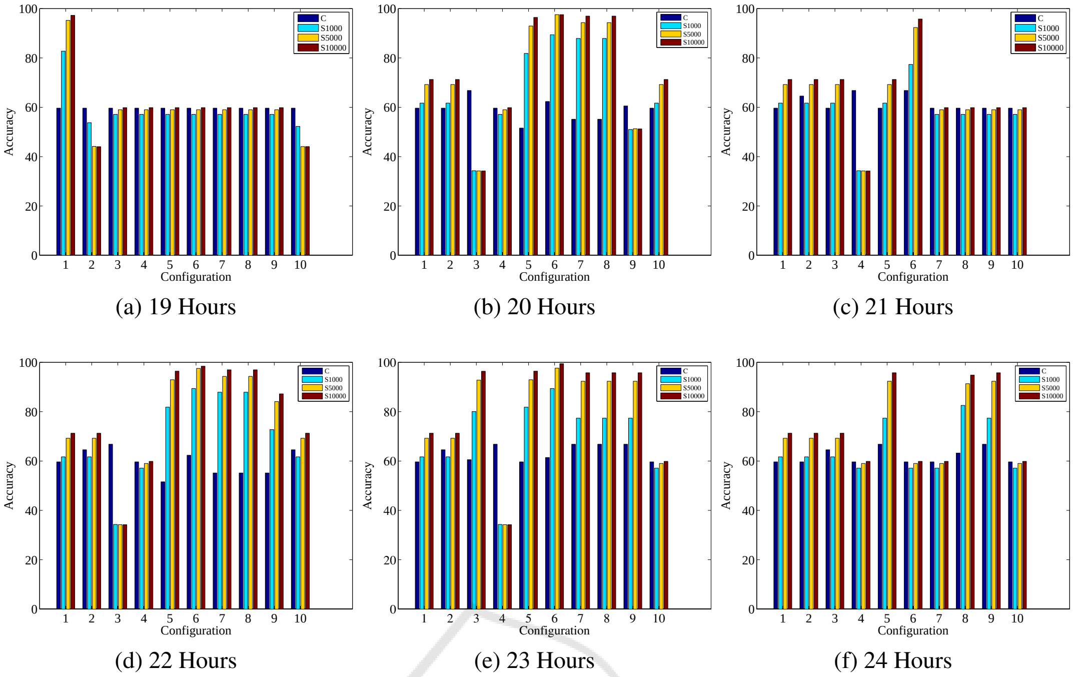
Figure 9: SIMO Accuracy Results 13-18.



(a) 13 Hours (b) 14 Hours (c) 15 Hours

(d) 16 Hours (e) 17 Hours (f) 18 Hours

Figure 10: SIMO FMeasure Results 13-18.



(a) 19 Hours

(b) 20 Hours

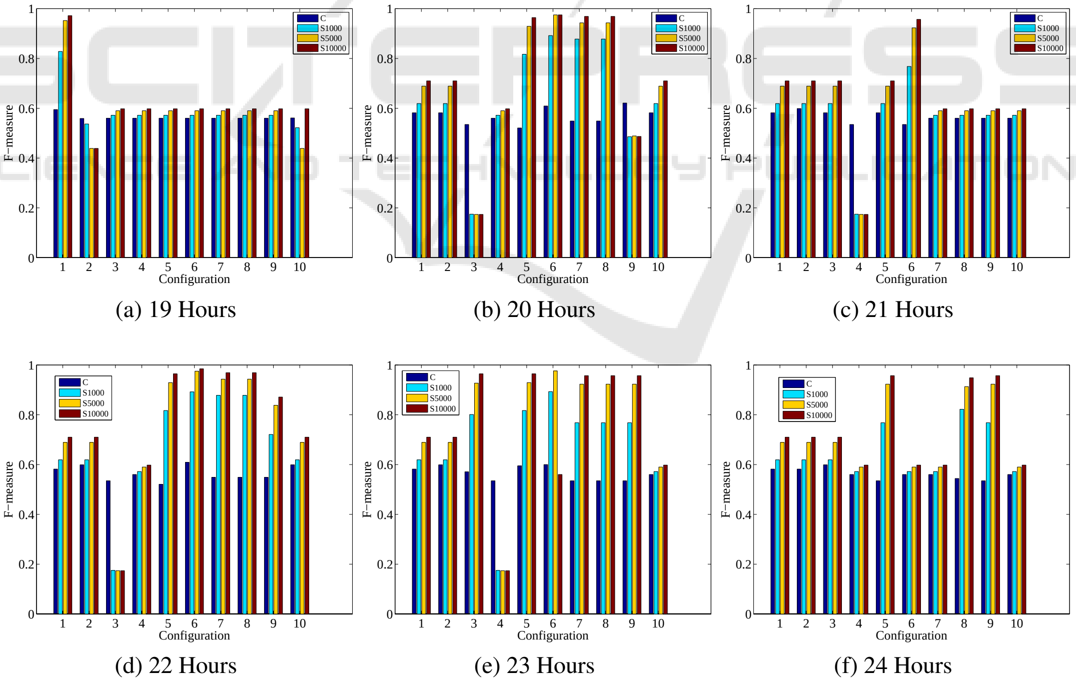
(c) 21 Hours

(d) 22 Hours

(e) 23 Hours

(f) 24 Hours

Figure 11: SIMO Accuracy Results 19-24.



(a) 19 Hours

(b) 20 Hours

(c) 21 Hours

(d) 22 Hours

(e) 23 Hours

(f) 24 Hours

Figure 12: SIMO FMeasure Results 19-24.