Representing Uncertain Time Series by Using Enhanced Symbolic Aggregate Approximation

Nabilah Filzah Mohd Radzuan, Zalinda Othman, Azuraliza Abu Bakar & Mohd Zakree Ahmad Nazri

Data Mining Lab, Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia

ABSTRACT

Abstract— Uncertain time series data is characterized by its numerical and continuous values. Knowledge from uncertain time series data brings important meaning for future prediction. As the uncertain time series datasets grow more dominant in a wide variety of applications, including weather, manufacturing, environmental, finance, economic, and medicine, the need for prediction without losing information and knowledge are high. An appropriate representation of uncertain time series data is required for more accurate prediction. Data representation is one of the most important tasks in time series data pre-processing. Prior to the representation, these uncertain data properties are being compressed without losing any valuable information. The compressed properties of the data are important in order to simplify the dataset for next data handling. Therefore, this study aims to propose an Enhanced Symbolic Aggregate Approximation (EN-SAX) as the basis for uncertain time series data representation. The experimental results show that the EN-SAX manages to represent the data with lower error rates. It also improves the prediction accuracy. This work will benefit in many application domains in terms of representing the uncertain datasets without losing valuable information.

Keywords: Uncertain time series; representation; EN-SAX.

INTRODUCTION

Time series mining is one of temporal data mining applications. Time series is well-known as a stretch of values on a similar scale, indexed by a time that occurs naturally in many application domains such as weather, manufacturing, environmental, economic, finance, and medicine. Uncertainty exists in time series. Uncertainty is a basic feature of automatic and semi-automatic data processes (Keijzer et al., 2007). There are many solutions have been proposed in order to reduce uncertainty because of risks in losing information and misleading results (Radzuan et al., 2013). Besides, the uncertain time series is also a non-negative and precisely different ways in some fields. Particularly, uncertain data refers to data in which the ambiguity on whether it takes place or not exists or data for that the attribute values are not ascertained with 100 percent probability (Hooshsadat & Za, 2012).

The combination of uncertainties is significant (Cloke & Pappenberger, 2009; Jankovic, 2004; Lykoudis et al., 2010) in time series and brings important knowledge for end user. In order to gain benefits through uncertain time series data, the essential problem of uncertainty should be focused. The essential problem in the circumstance of time series data mining is how to represent the uncertain time series data. Data representation is one of the most important tasks in time series data pre-processing. Prior to the representation, these uncertain data properties are being compressed without losing any valuable information. The compressed properties of the data are important to simplify the dataset for next data handling. Therefore, this study aims to propose an Enhanced Symbolic Aggregate Approximation (EN-SAX) which uses mean values as the basis for uncertain time series data representation.

The structure of this paper is organized as follows. After the introduction, the main aims of the paper and briefly related research methods will be defined and described in part 2. The experiment on uncertain time series through Enhanced Symbolic Aggregate Approximation (EN-SAX) method will be explained in part 3, including its challenges and results. In part 4, an extensive discussion on the objectives of the investigated research, their approaches, and applied datasets are provided. The final sections of the paper contain the conclusions and references.

Related work

Uncertainty has been explicitly indicated as one of the future challenges in many fields (Halevy & Ordille, 2006). There is relationship between original series (certain) and uncertain series of data in time series. The certain time series data is a proper time series which has been corrected, or the inaccurate records have been removed from the dataset (Zuo et al., 2011). Thus, the certain time series is extracted to represent the original uncertain time series data. Prior to the representation, these uncertain time series data is being processed for valuable information. Representing time series data can be worked in various methods(Barnaghi et al., 2013).

2.1 Piecewise Aggregate Approximation

Piecewise Aggregate Approximation (PAA) presents the time series data using the mean (Avg) value of each segment (Fu, 2011) is the simplest method of representation. PAA using length of a time series and size after dimensionality reduction (Keogh & Pazzani, 2000; Yi et al., 2000) and it originally called piecewise constant approximation (Buu & Anh, 2011; Keogh & Pazzani, 2000). This method can be used although the rate of sampling is too low and missing some values is not important.

2.2 Adaptive Piecewise Constant Approximation

In 2003, (Lin et al., 2003) is proposed an extended version of representation named Adaptive Piecewise Constant Approximation (APCA). This method process with the length of each segment is not fixed with high possibility to prominent loss patterns in different segments.

2.3 Symbolic Aggregate Approximation

The common method for representing time series data is Symbolic Aggregate Approximation (SAX). SAX transfer numeric time series to a new form of data. Then, SAX discretises the time series data into segments and transforms of each segment into the symbol (Ahmed et al., 2010; Aref et al., 2004).

2.4 Extended Symbolic Aggregate Approximation

Lkhagva introduced a new time series data representation, i.e. Extended SAX (ESAX) for financial applications (Lkhagva & Suzuki, 2006). This ESAX uses Min, Avg, Max values as a string of symbols to search similarity of shapes between financial time series. But ESAX method has no numerical values as output to measure the similarity between original data and represented data (Lkhagva & Kawagoe, 2006; Lkhagva & Suzuki, 2006).

2.5 Enhanced Symbolic Aggregate Approximation

Although, PAA, APCA, SAX, and ESAX are appropriate methods for time series dimensionality reduction, but these methods are not suitable for uncertain time series data. This is because of the methods are based on mean values estimate, and there are some patterns with the high possibility to loss some of this kind of patterns. Hence, Enhanced SAX (EN-SAX) is chosen to be used for uncertain time series data. The EN-SAX are using Min, Avg, Max as a vector of data for data similarity search, using K-Means clustering method to determine symbols zones, and using cosine similarity to determine similarity between vectors of data (Barnaghi et al., 2013). Therefore, uncertain time series data can be experimented through this method.

EN-SAX is based on SAX and it uses two additional Min and Max points with original mean values of each segment in time series data. This helps to preserve some important points that are meaningful especially in uncertain time series. The Min and Max points in EN-SAX help to detect important values and improve the accuracy. There are four main steps in implementing this representation method for time series data. The steps are included (i) representation and indexing, (ii) similarity measurement, (iii) segmentation and (iv) pattern discovery.

2.5.1 Representation and indexing

Detection of uncertainty in time series data is done before implementing clustering. Prior to detection, existence of uncertainty in time series dataset is proved first. The highest percentage of uncertainty represents the highest uncertainness in the data (Radzuan et al., 2014). The purpose of this is to prove that the dataset has accomplished the experiment requirement. The clustering is being implemented where the current data is grouped into different clusters. During this step, in order to implement EN-SAX method, it is needed to map each set of data to a special symbol. Table 1 shows the example of mapping of each cluster to a symbol.

Table 1
Mapping clusters to symbols

Cluster	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Symbol	A	B	C	D	E	F	G	H	I	J
Cluster	10	11	12	13	14	15	16	17	18	19
Symbol	K	L	M	N	O	P	Q	R	S	T

2.5.2 Similarity measurement

The original exchange data is transferred to new symbolized form. The description of evaluation of error rate between original data and the new symbolized data is done. Then, the mean value for each symbol using the number of the records for each symbol is calculated as shown in Figure 1.

Min (A) = (average[cell(A1…Ax,1)]) (1)
//Min value column shows with 1

Avg (A) = (average[cell(A1…Ax,2)]) (2)
//Min value column shows with 2

Max (A) = (average[cell(A1…Ax,3)]) (3)
//Min value column shows with 3

Mean (A) = [Min (A), Avg (A), Max (A)] (4)

Figure 1 The mean (average) value provides a vector with three values [Min, Avg, Max]

The error rate can be calculated by comparing each segmented data that is mapped to a symbol with equal symbol mean value. As example, value for segment 1 of data are shown with S1=[Min, Avg, Max] that corresponds to a cluster symbol such as A as shown in Figure 2 and calculation of Min, Avg and Max value for each symbol shown in Figure 3.

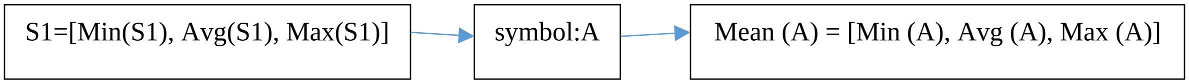

S1=[Min(S1), Avg(S1), Max(S1)]
symbol:A
Mean (A) = [Min (A), Avg (A), Max (A)]

Figure 2 Calculate error rate

Symbol	Min	Avg	Max
A_1	$A_{1,1}$	$A_{1,2}$	$A_{1,3}$
…	…	…	…
A_x	$A_{x,1}$	$A_{x,2}$	$A_{x,3}$

Symbol	Min	Avg	Max
A	Min(A)	Avg(A)	Max(A)

Figure 3 Calculation of Min, Avg and Max value for each symbol

2.5.3 Segmentation

During data segmentation process, the data must be divided to equal sized parts or called as segment. The size of the segment is determined by type of data and application domain. The mean value for each segment is calculated and used for representation of data for each segment. The single value for each segment is represented in SAX will then change to a vector in EN-SAX. The vector represented in EN-SAX is represented with [Min, Avg, Max] value with another attribute that is used as id for each segment.

2.5.4 Pattern discovery

The final step is categorizing and grouping different cases on the available data. WEKA is chosen as data mining tool for clustering method. The changed data to segments are clustered in order to represent with three values for each segment. The parameters for clustering are determined based on the data type and application domain.

Experiments

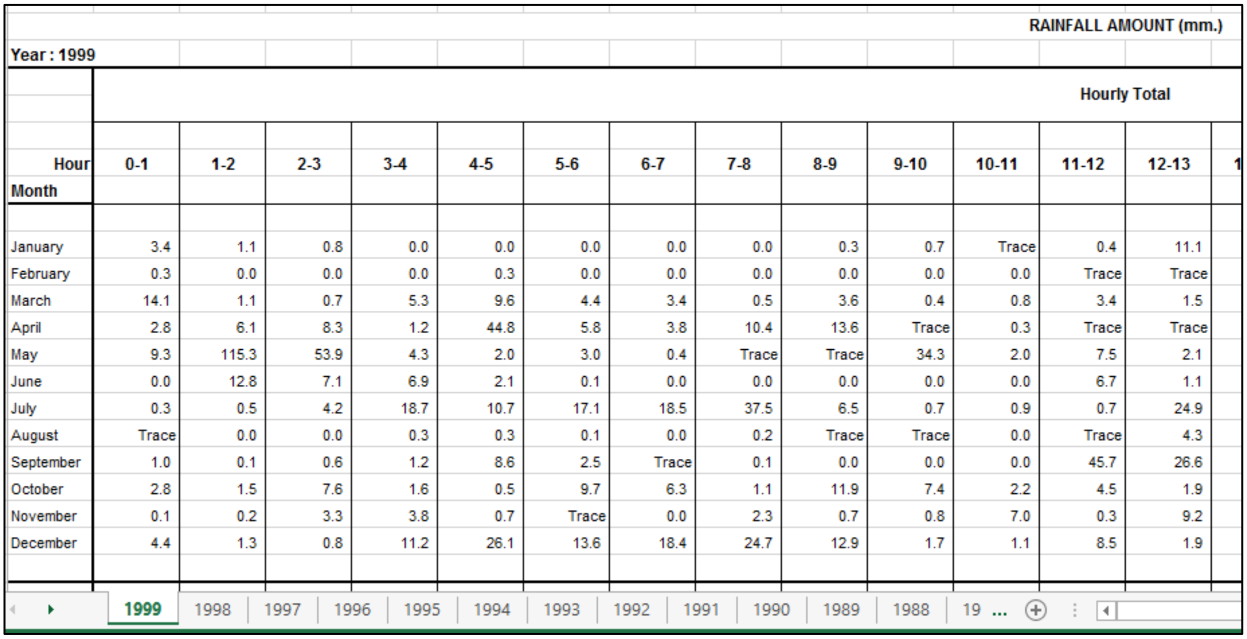
The experiment in this study focuses on proposing the EN-SAX, which uses mean values as the basis for uncertain time series data representation.The uncertain data is used to prove especially the accuracy of each prediction so that these methods can be studied for time series data. The performance of EN-SAX is evaluated in terms of error rate and prediction accuracy. It is also compared with the original SAX method in order to prove the possibility to be loss of important patterns. First, the data is prepared for EN-SAX and then the data will go through Linear Regression (LR) and Support Vector Machines (SVM) method to perform the prediction.

3.1 Data collection and preparation

The real rain dataset that is time series datasets from Petaling Jaya Station is used for the experiment. The rainfall dataset of 20 years from year 1980 to 1999 is shown in Figure 4. The detection of uncertainty in time series dataset is been done first before proceeding to representation process as mention in previous section. Particularly, to add the 10 percent uncertainty to an attribute, it is attached with a 0.9 probability and the remaining 0.1 is distributed randomly among other values appear in the domain (Hooshsadat & Za, 2012). Eventually, the highest percentage of uncertainty represents the highest uncertainness in the data (Radzuan et al., 2014). This process helps in avoiding loss of information in the dataset compared with a normal time series process.

The rainfall data is chosen as a case study for this experiment. The attributes are including months from January to December, for 30 days and in 24 hours per day. There are trace value in rainfall dataset that represent value less than 0.1 mm. The data is saved as the main dataset in MS Excel sheets before applying various steps in this study. The 20 years' time series data are divided into Dataset 1 (DS1), Dataset 2 (DS2), Dataset 3 (DS3), Dataset 4 (DS4), and Dataset 5 (DS5).

The rainfall dataset is divided into equal size of parts that are defined as segments. Using the values included in each segment, the minimum (Min), mean (Avg), and maximum (Max) value is calculated for each segment. These values are calculated in order to implement the EN-SAX method as representation solution for uncertain time series data. Therefore, in the first step of representation, Min, Avg and Max value are calculated for each segment. It is needed to map each set of data to a special symbol before implement in EN-SAX method. Each segment is mapped to related symbol based on Min, Avg and Max values. The Avg value indicates the domain of each cluster which maps to the related symbol.



RAINFALL AMOUNT (mm.)

Year : 1999

Hourly Total

Hour Month	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13
January	3.4	1.1	0.8	0.0	0.0	0.0	0.0	0.0	0.3	0.7	Trace	0.4	11.1
February	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	Trace	Trace
March	14.1	1.1	0.7	5.3	9.6	4.4	3.4	0.5	3.6	0.4	0.8	3.4	1.5
April	2.8	6.1	8.3	1.2	44.8	5.8	3.8	10.4	13.6	Trace	0.3	Trace	Trace
May	9.3	115.3	53.9	4.3	2.0	3.0	0.4	Trace	Trace	34.3	2.0	7.5	2.1
June	0.0	12.8	7.1	6.9	2.1	0.1	0.0	0.0	0.0	0.0	0.0	6.7	1.1
July	0.3	0.5	4.2	18.7	10.7	17.1	18.5	37.5	6.5	0.7	0.9	0.7	24.9
August	Trace	0.0	0.0	0.3	0.3	0.1	0.0	0.2	Trace	Trace	0.0	Trace	4.3
September	1.0	0.1	0.6	1.2	8.6	2.5	Trace	0.1	0.0	0.0	0.0	45.7	26.6
October	2.8	1.5	7.6	1.6	0.5	9.7	6.3	1.1	11.9	7.4	2.2	4.5	1.9
November	0.1	0.2	3.3	3.8	0.7	Trace	0.0	2.3	0.7	0.8	7.0	0.3	9.2
December	4.4	1.3	0.8	11.2	26.1	13.6	18.4	24.7	12.9	1.7	1.1	8.5	1.9

Figure 4 Rainfall dataset at Petaling Jaya Station

3.2 Similarity measurement

The SAX method, as a common method for time series data representation, is used on uncertain time series data where the original exchange data is transferred to a new symbolic form through EX-SAX method. The measurement of similarity starts between original data and the new symbolized data by calculating the error rate between the two sets of original and new data. The mean value is used in calculating the error rate. Table 2 shows average error rate between original and symbolized data using EN-SAX. The comparison results between SAX and EN-SAX show a decrease in error rate values between original data with segmented data. Table 3 shows the result of the experiment

Table 2 *Average error rate between original and symbolized data using EN-SAX*

Symbol	DS1	DS2	DS3	DS4	DS5
A	1.4	3.3	11.9	0.6	0.1
B	0.4	1.3	0.2	6.6	8.7
C	0.3	0.1	0.2	0.2	0.0
D	7.1	5.7	3.6	7.3	12.0
E	2.8	0.2	1.0	0.1	12.8
F	1.4	3.5	9.6	1.3	1.5

Table 3 *The result of experiment for comparison results between SAX and EN-SAX which decrease in error rate values between original data with segmented data*

Dataset	SAX	EN-SAX
DS1	0.64	0.60
DS2	0.57	0.55
DS3	0.68	0.64
DS4	0.71	0.69
DS5	0.67	0.61

3.3 Prediction error rate on discretized data

The data is prepared using SAX and EN-SAX that is read from related files and applying LR and SVM methods for prediction. The results in Table 4 shows the comparison between performance of LR and SVM method where LR has better performance for prediction as shown by prediction error rate. The less error rate in dataset, the better prediction can be made.

Table 4 *The comparison between performance of LR and SVM method*

Dataset	LR		SVM	
	SAX	EN-SAX	SAX	EN-SAX
DS1	0.74	0.54	0.69	0.50
DS2	0.67	0.47	0.70	0.45
DS3	0.77	0.58	0.81	0.54
DS4	0.80	0.61	0.83	0.59
DS5	0.77	0.57	0.98	0.51

DISCUSSION

The yield of representation of uncertain time series data brings important meaning for future prediction. In this study, the results from the experiment show that uncertain time series data can be represented using EN-SAX. The performed analytical and experimental of this study towards EN-SAX has proved that EN-SAX was suitable to represent uncertain time series data.

There are differences between uncertain data and uncertain time series data. While uncertain data refers to static data (Aggarwal et al., 2009), uncertain time series data refers to continuous data (Gagne et al., 2011). However, both collected data often inaccurate and are based on incomplete or inaccurate information that needed to be represented using the EN-SAX before proceeds to next steps.

As it is discussed in earlier study, EN-SAX method is more effective in comparison with other dimensionality reduction methods such as SAX. The features of EN-SAX include keeping important patterns on uncertain time series data and avoiding missing key values that may have significant role on the decision for future prediction. Therefore, the EN-SAX method is helpful in uncertain time series data.

CONCLUSION

This study proposes an improved method name Enhanced Symbolic Aggregate Approximation (EN-SAX) for uncertain time series data representation. The vector represented in EN-SAX is represented with [Min, Avg, Max] value with another attribute that is used as identification for each segment. The results show that EN-SAX method is more accurate and effective in using the represented data for prediction purposes in uncertain time series data. The results show that EN-SAX method able to represent uncertain time series data that produced more accurate and effective prediction model.

ACKNOWLEDGEMENTS

This work is acknowledged by Universiti Kebangsaan Malaysia (UKM), Malaysia through Research University (RU) grant scheme number of Project FRGS/1/2014/ICT02/UKM/02/1 and ERGS/1/2013/ICT07/UKM/02/4 for financial support.

REFERENCES

Aggarwal, C. C., Li, Y., Wang, J., & Wang, J. (2009). Frequent pattern mining with uncertain data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09* (p. 29). New York, New York, USA: ACM Press. doi:10.1145/1557019.1557030

Ahmed, A. M., Bakar, A. A., & Hamdan, A. R. (2010). Improved SAX time series data representation based on Relative Frequency and K-Nearest Neighbor Algorithm. In *2010 10th International Conference on Intelligent Systems Design and Applications* (pp. 1320–1325). IEEE. doi:10.1109/ISDA.2010.5687092

Aref, W. G., Elfeky, M. G., & Elmagarmid, A. K. (2004). Incremental, online, and merge mining of partial periodic patterns in time-series databases. *IEEE Transactions on Knowledge and Data Engineering*, *16*(3), 332–342. doi:10.1109/TKDE.2003.1262186

Barnaghi, P. M., Bakar, A. A., & Othman, Z. A. (2013). Enhanced symbolic aggregate approximation (EN-SAX) as an improved representation method for financial time series data. *International Journal of Soft Computing*, *8*(4), 261–268. doi:10.3923/ijscomp.2013.261.268

Buu, H. T. Q., & Anh, D. T. (2011). Time series discord discovery based on iSAX symbolic representation. *Proceedings - 2011 3rd International Conference on Knowledge and Systems Engineering, KSE 2011*, 11–18. doi:10.1109/KSE.2011.11

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3-4), 613–626. doi:10.1016/j.jhydrol.2009.06.005

Fu, T. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181. doi:10.1016/j.engappai.2010.09.007

Gagne, D. J., McGovern, A., & Xue, M. (2011). Machine learning enhancement of storm scale ensemble precipitation forecasts. In *Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation - KDMS '11* (p. 45). New York, New York, USA: ACM Press. doi:10.1145/2023568.2023581

Halevy, A., & Ordille, J. (2006). Data Integration : The Teenage Years. In *VLDB '06 Proceedings of the 32nd international conference on Very large data bases* (Vol. 12, pp. 9–16).

Hooshsadat, M., & Za, O. R. (2012). An Associative Classifier For Uncertain Datasets. In *Advances in Knowledge Discovery and Data Mining* (p. pp 342–353). doi:10.1007/978-3-642-30217-6_29

Jankovic, V. (2004). Science Migrations: Mesoscale Weather Prediction from Belgrade to Washington, 1970–2000. *Social Studies of Science*, *34*(1), 45–75. doi:10.1177/0306312704040490

Keijzer, D. A., Keulen, V. M., & Dekhtyar, A. (2007). Report on the First VLDB Workshop on Management of Uncertain Data (MUD). *ACM SIGMOD Record*, *36*(4), 18–32. doi:10.1145/1361348.1361363

Keogh, E. J., & Pazzani, M. J. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases. *Knowledge Discovery and Data Mining, Proceedings*, *1805*, 122–133. doi:10.1007/3-540-45571-X_14

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD '03* (p. 2). New York, New York, USA: ACM Press. doi:10.1145/882085.882086

Lkhagva, B., & Kawagoe, K. (2006). New Time Series Data Representation ESAX for Financial Applications. *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, x115–x115.

doi:10.1109/ICDEW.2006.99

Lkhagva, B., & Suzuki, Y. (2006). Extended sax: Extension of symbolic aggregate approximation for financial time series data representation. *DEWS2006 4A-i8*, *7*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.9325&amp;rep=rep1&amp;type=pdf\nhttp://www.ieice.or.jp/iss/de/DEWS/DEWS2006/doc/4A-i8.pdf

Lykoudis, P. S., Argiriou, A. A., & Dotsika, E. (2010). Spatially interpolated time series of δ18O in Eastern Mediterranean precipitation. *Global and Planetary Change*, *71*(3-4), 150–159. doi:10.1016/j.gloplacha.2009.09.004

Radzuan, N. F. M., Othman, Z., & Bakar, A. A. (2013). Uncertain Time Series in Weather Prediction. *Procedia Technology*, *11*(Iceei), 557–564. doi:10.1016/j.protcy.2013.12.228

Radzuan, N. F. M., Othman, Z., & Bakar, A. A. (2014). Analysis of Uncertainty in Time Series Data: Issues and Challenges. In *The Asian Conference on Technology, Information & Society 2014* (pp. 13–24).

Yi, B, K., & Faloutsos, C. (2000). Fast Time Sequence Indexing for Arbitrary Lp norms. *Proc. of the 26st Int. Conf. on VLDB'00*, 385–394. doi:10.1016/j.cppeds.2011.05.001

Zuo, Y., Liu, G., Yue, X., Wang, W., & Wu, H. (2011). Similarity Matching over Uncertain Time Series. *2011 Seventh International Conference on Computational Intelligence and Security*, 1357–1361. doi:10.1109/CIS.2011.302