Analysis of Uncertainty in Time Series Data: Issues and Challenges

Nabilah Filzah Mohd. Radzuan, University Kebangsaan Malaysia, Malaysia
Zalinda Othman, University Kebangsaan Malaysia, Malaysia
Azuraliza Abu Bakar, University Kebangsaan Malaysia, Malaysia

The Asian Conference on Technology, Information and Society 2014
Official Conference Proceedings

Abstract
This paper reviews issues and challenges of uncertainty in time series data. The aim of uncertainty analysis is to determine the ways of how to deal with uncertain data in order to gain knowledge, fit low dimensional model, and do prediction. So as to build an efficient predictive tool, uncertainty in data could not be ruled out because it may bring important knowledge. Uncertainty information arises from different resources such as process uncertainty, model uncertainty or data uncertainty. In this paper, issues and challenges of these uncertainties in time series data will be discovered and how these issues could be solved by data mining techniques will be discussed. Frequent pattern mining algorithm through FP-growth, Apriori algorithm and H-mine are methods that could be used to investigate the existing of uncertainty data. Meanwhile, Euclidean distance, particle swarm optimization, Monte Carlo simulation, and regression are methods that could be compared as prediction methods. These methods have been implemented in many data types since early 1900s. Also, this paper shows results of the uncertainty detection test on time series data sets. The test aims to prove the existing of uncertainty in the data. This work will benefit in many application domains.

Keywords: uncertainty, uncertain time series, mining algorithm

Introduction

Uncertainty is a basic feature of automatic and semi-automatic data processes (Keijzer, Keulen, & Dekhtyar, 2007). There are many solutions have been proposed to reduce uncertainty because of risks in losing relevant information and misleading results (Radzuan, Othman, & Bakar, 2013). Uncertainty exists in time series data. Time series data is known as a stretch of values on a similar scale, indexed by a time that occurs naturally in many application domains such as environmental, economic, finance, and medicine. The aim of time series analysis is to formulate time series data to gain knowledge, to fit low dimensional models, and to make predictions. In reality, time series also deals with uncertainty. An uncertain time series data is a non-negative and precisely different ways in a number of fields (Cloke & Pappenberger, 2009; Lykoudis, Argiriou, & Dotsika, 2010). Particularly, uncertain data refers to data in which the ambiguity on whether it really 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) and brings important knowledge.

However, there are challenges to deal with when it involves domains such as manufacturing and weather forecast. Among the challenges include limited observational basis for seasonal and long term prediction, accurate forecasting of weather that may poses danger to aviation, prediction of product yielded during production process when expected situations happen, and many more (Williams et al., 2008). Researchers in manufacturing have attempted to discover the appropriate techniques specifically for modelling and processing uncertain time series for temporal data (Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012a; Dallachiesa, 2011). The involvement of this modelling and processing for uncertain time series is significant because they deal with query efficiency for accurate results (Zuo, Liu, Yue, Wang, & Wu, 2011). Meanwhile, in weather forecasting, new discovered knowledge from uncertain time series could be used in weather prediction or precipitation for the benefits of human being.

Since 1900s, there have been many studies carried out in dealing with uncertain time series. However, they used different terms. In 1905, researchers determined the relationships between normal and abnormal embryo development of a frog (Morgan, 1905). The inconsistency in the segmentation process was regarded as uncertain time series in the development stages (Hume, 1911). Meanwhile, in 1985, researchers helped identifying wet spells of weather and assessed the unusualness of the recent episode of heavy precipitation in meteorology department even though the uncertain whether prolonged the dry spell, which stroke the lake levels down to much lower levels before the onset of the next severe wet spell (Karl & Young, 1985). These situations visualize that the uncertain time series is very helpful and useful when knowledge is extracted from the data sets. Consequently, it is a promising direction to explore for more knowledge extraction methods in uncertain time series mining.

Further, uncertain time series mining is believed to be able to avoid risks and help in making better daily decisions. Uncertain time series mining also can improve the quality of demand, and identify temporal patterns that emerge and persist. This paper briefly describes the analysis of uncertain time series through issues and challenges. In response to that, this paper showcases the existence of uncertainty in selected uncertain time series datasets. In the remaining parts of this paper, some related works are discussed in Sec. 2. Then, the analysis results are exhibited in Sec. 3. Next, Sec. 4 discusses the techniques and benefits. Finally, the conclusion is drawn in Sec.5.

Related Work

Uncertainty has been explicitly indicated as one of the future challenges in many fields (Halevy & Ordille, 2006). The uncertainty presents in all data processes and methods whether realize or not. The characterizing feature of uncertainty and other early works using uncertainty theories is after probabilities is used to select matching and non-matching objects (Hayne & Ram, 1990). Therefore, the uncertainty generated during data processes is lost (Dey, Sarkar, & Society, 2002; Florescu, Koller, & Levy, 1997).

There are relationships between original series (certain) and uncertain (Haitao & Xiaofu, 2009; Russa & Andrews, 2010). Generally, a clean time series data (certain time series data) is chosen for experiment. A certain time series data is a proper time series data which has been corrected or the inaccurate records have been removed from dataset. The certain time series is extracted to represent the original uncertain time series (Zuo et al., 2011). Uncertain time series can be treated as positional uncertain vectors (Abfalg, Kriegel, Kr, & Renz, 2009).

Besides, uncertainty exists in a modelling process, in which it arises from a fundamental choice as seen in grid resolution and from the parameterization of unresolved processes at the grid scale (Angew et al., 2004). Also, a high uncertainty brings big impact on prediction of regional climate change (Hawkins & Sutton, 2009). In fact, a lack of model diversity can cause a limited range of projections in climate change (Pennell & Reichler, 2011). Meanwhile, the distinct sources of uncertainty in prediction include internal variability, model uncertainty or response uncertainty, and scenario uncertainty (Hargreaves, 2010; Hawkins & Sutton, 2009).

The uncertain time series has been explored extensively in recent years. Uncertainty can be due to data aggregation, privacy-preserving transforms, and error-prone mining algorithms (Dallachiesa et al., 2012a; Dallachiesa, 2011). As a result, they found that uncertainty information might appear on different reasons (Dallachiesa et al., 2012a). Predicting uncertain time series appears to be a serious problem, as the existing forecast of certain time series does not purely mirror the ability of predicting future decisions. Uncertain time series in prediction is believed can avoid risks and help in making better daily decisions.

As an example, uncertain data is created by several applications in data forecasting, as can be seen in weather precipitation predicting for meteorology department, or in manufacturing demand prediction, which both actually can gain benefits in handling future outcome. Uncertain time series is important in making predictions. It influences the changeable climate that provides more useful information. Then, important knowledge can be tackled from this changeable gap that exists in uncertain time series data, in which the uncertainty can provide better results in terms of quality and efficiency (Dallachiesa et al., 2012a; Dallachiesa, 2011; Zuo et al., 2011).

Hence, the determination of predicting uncertain time series should be noted as a serious action to improve the quality of yield. The limitation found from the analysis can be used as an opening of the experiment and aim for securing the limitation for enhancing the prediction outcome. Previous studies have discovered some possible properties of uncertainty in dataset (see Table 1). Also, clarification of uncertainty in dataset is important in identifying the type of data, so that they are not simply neglected. In normal practice, the organizer will neglect any data that they perceive as 'error' without investigating uncertain data's properties.

Table 1: The Properties of Uncertainty in Dataset

The properties
• non-negative • loss value or null, truly different ways in a number of fields • data aggregation • privacy-preserving transforms • error-prone mining • positional uncertain vectors • exist in the modelling process where it arises from fundamental choice • and, from the parameterization of processes unsolved at grid scale. (Abfalg et al., 2009; Angew et al., 2004; Cloke & Pappenberger, 2009; Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012b; Dallachiesa, 2011; Lykoudis et al., 2010)

Therefore, there is an initiative to implement a number of algorithms consecutively to detect the uncertainty in dataset. In regards to that, Uncertain Associative Classifier (UAC) method (Hooshsadat & Za, 2012) could be used. It is measured partly on its accuracy, in which the percentage of accuracy is calculated using rule-based classifier on datasets. It is modelled based on a direct mining of discriminative patterns for classifying uncertain data at the level of uncertainty. In conjunction to this, a previous study found that the accuracy of data can be determined by uHARMONY, DTU, and uRule methods through UCI datasets (Hooshsadat & Za, 2012). Then again, uncertainty information arises from different resources such as process uncertainty, model uncertainty or data uncertainty. Frequent pattern mining algorithms through FP-growth, Apriori algorithm and H-mine are methods that could be used to investigate the existing of uncertainty in data. Table 2 shown the approaches include the advantages and disadvantages.

Table 2: The Uncertain Data Approaches

Apriori (UApriori)	FP-growth (FP-tree)	H-mine (UH-mine)
Apriori identify the frequent items in the database and extending them to larger item sets appear sufficiently often in the dataset. - UApriori is an extended from Apriori Algorithm. - Efficient by employing pruning method.	Efficient and scalable especially for dense dataset - Loss of compression properties. - Large number of false positive is generated. - The elimination of dataset further affects the efficiency.	Efficient and scalable especially for uncertain dataset. - Can avoid generating a large number of candidate itemsets. - Reduce memory requirements. - Best trade-off in terms of running time and memory usage.

Experiment and Result
The experiment in this study focuses on identifying uncertainty in selected datasets. 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 data has gone through a discretization process (a process of organizing the dataset in minimizing redundancy and dependency, and makes it more informative to use).The discretization process involves scale-selective discretization (SSD) procedure as in (Vuorinen, Larmi, Schlatter, Fuchs, & Boersma, 2012). This SSD separates small and large scales of the flow using a high-pass filter.

The Apriori is used as a generate-and-test approach by generating the dataset attributes and testing if they are frequent or not. Generation of dataset attributes are disconnected, where it is involve checking subset in each attributes and scanning multiple databases. Then, FP-Growth allows frequent attributes discovery without dataset attributes generation. There are two steps in this approach; first, it builds a compact data structure called FP-tree where it is built using two passes over the dataset. Second, it extracts frequent attributes directly from FP-tree where traversal through FP-tree. The H-mine tries to avoid generating a large number of dataset attributes and uses all involved attributes without eliminating or avoiding the null value.

The three approaches are intersect with Uncertain Associative Classifier (UAC) method as implemented in (Hooshsadat & Za, 2012). The UAC algorithm can only be implemented after the trained dataset goes through a discretization process. The UAC algorithm is visualized in Figure 1. It involves three stages of UAC rule filtering of the three approaches. Further, the algorithm of each stage is detailed in Appendix A. Briefly, the UAC algorithm selects one classifying rule for each instance which has the highest relative precedence with respect to the test instance (Hooshsadat & Za, 2012).

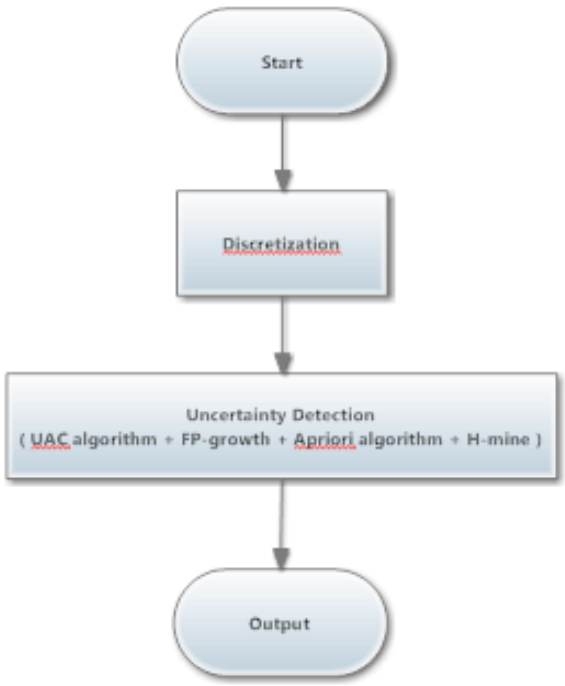

Figure 1: Flow of approaches

In this experiment, 43 uncertain time series datasets from UCI benchmark data were used. The results in Table 2 show the existence of uncertainty in the datasets. 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.

Table 2 Result of uncertainty percentage detection for 43 dataset

Dataset	Uncertainty (%)
Single Chest Mounted	68.2
ADLs	24.9
Amazon Access	70.4
ASL	44.3
ASL signs	39.3
Bach Chorales	83.6
Buzz	28.1
CalIt2	32.0
Character Trajectories	85.7
DS Activities	41.9
Daphnet	37.3
Wrist worn	21.5
Diabetes	12.4
EEG	20.0
EEG Eye State	25.5
EMG Lower	88.1
GSA Drift	0
GSA turbulent gas mixtures	23.8
GSA under flow modulation	48.0
GSA open sampling settings	26.0
Gesture Phase Segmentation	31.8
Smartphones	34.0
ICU	50.4
Individual household	10.8
Istanbul SE	39.8
Japanese Vowels	11.1
Localization Data	87.9
Opportunity Activity	42.3
Ozone	14.3
PAMAP2	39.3
PEMS-SF	23.6
Pioneer-1	28.1
Predict keywords	32.0
Pseudo Periodic Synthetic	85.7
Realdisp Activity	61.9
Robot	77.3
SML2010	81.5
Spoken Arabic Digit	12.4
Synthetic Control	80.0
URL Reputation	23.0
Walking Activity	34.3
Vicon	75.0

The experiment explained in the previous paragraph proves the existence of uncertainty in the time series dataset. The time series dataset was normalized before implemented on UAC algorithm. Normalization is important in order to minimize redundancy and isolate data. The mining process of time series data differs from normal dataset as the data properties itself are different. The result of time series dataset is same with previous study (Hooshsadat & Za, 2012) where there is existence of uncertainty in the dataset.

Discussion
The yield of prediction and knowledge from uncertain data brings important meaning for future prediction especially in weather domain. In a real situation, unpredictable events happen without being anticipated. In this study, the reviewed methods bring benefit to domain in predicting the uncertain time series data. The performed analytical and experimental comparisons of techniques described in the previous section should be further experimented in order to get accurate prediction.

The FP-growth approach extracts frequent attributes from the FP-tree. The FP-tree can be built if only consider the transactions containing a particular attributes or else removing the attributes from all transactions. The H-mines approach help in minimizing items lost from that transactions. The compressed datasets have high tendency of losing attributes. All the uncertain properties in the datasets have been calculated and shown in percentage. From the percentage values, the uncertainty in datasets are detected.

In this study, there are differences between uncertain data and uncertain time series data. While uncertain data refers to static data (Aggarwal, Li, Wang, & Wang, 2009), uncertain time series data refers to continuous data (Gagne, McGovern, & Xue, 2011). However, both collected data often inaccurate and are based on incomplete or inaccurate information. The detection test on uncertainty has shown that there are uncertainties in the datasets that would bring highly potential in yielding information for future prediction. The test helps the organizer to not neglect any data that they perceive as ‘error’. Therefore, the UAC method could be utilized for time series data in determining uncertainty in data. Then, the yield, which is uncertain time series data, of the process can be implemented on prediction methods.

Conclusion
This paper explains on evaluation methods used in uncertain time series. The analysis on previous works and the experiments outline the methods on certain data in order to extract knowledge for future work. This study discovers that there are methods that bring limitation in their prediction processes. The presence of uncertainty in dataset can be determined through a combination of FP-growth, Apriori algorithm, H-mine and UAC method. The data first gone through a discretization process involving SSD, in which it is a process of organizing the dataset in minimizing redundancy and dependency, and make it more informative to use. Through the experiment, on uncertainty existence in uncertain datasets have proved that uncertainties exist in time series data. Although the experiment brings benefits to domain, still future actions must be taken in obtaining accurate prediction in uncertain time series.

References

Abfalg, J., Kriegel, H., Kr, P., & Renz, M. (2009). Probabilistic Similarity Search for Uncertain Time Series. In *SSDBM 2009 Proceedings of the 21st International Conference on Scientific and Statistical Database Management* (pp. 435 – 443). doi:10.1007/978-3-642-02279-1_31

Aggarwal, C. C., Li, Y., Wang, J., & Wang, J. (2009). FREQUENT PATTERN MINING ALGORITHMS WITH UNCERTAIN DATA. In *Managing and Mining Uncertain Data* (pp. 428–459).

Angew, M. R. S., Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., … Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature Publishing Group*, *430*(August 2004), 768–772. doi:10.1038/nature02770.1.

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

Dallachiesa, M. (2011). Similarity Matching for Uncertain Time Series : Analytical and Experimental Comparison, 8–15.

Dallachiesa, M., Nushi, B., Mirylenka, K., & Palpanas, T. (2012a). Uncertain Time-Series Similarity : Return to the Basics. In *Proceedings of the VLDB Endowment* (Vol. 5, pp. 1662–1673).

Dallachiesa, M., Nushi, B., Mirylenka, K., & Palpanas, T. (2012b). Uncertain Time-Series Similarity : Return to the Basics, *5*(11), 1662–1673.

Dey, D., Sarkar, S., & Society, I. C. (2002). Generalized Normal Forms for Probabilistic Relational Data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *14*(3), 485–497.

Florescu, D., Koller, D., & Levy, A. Y. (1997). Using probabilitistic Information in Data Integration. In *Proceeding VLDB '97 Proceedings of the 23rd International Conference on Very Large Data Bases* (pp. 216–225).

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

Haitao, L., & Xiaofu, Z. (2009). Precipitation Time Series Predicting of the Chaotic Characters Using Support Vector Machines. In *2009 International Conference on Information Management, Innovation Management and Industrial Engineering* (pp. 407–410). Ieee. doi:10.1109/ICIII.2009.407

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).

Hargreaves, J. C. (2010). Skill and uncertainty in climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(4), 556–564. doi:10.1002/wcc.58

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1107. doi:10.1175/2009BAMS2607.1

Hayne, S., & Ram, S. (1990). Multi-user view integration system (MUVIS): an expert system for view integration. In *[1990] Proceedings. Sixth International Conference on Data Engineering* (pp. 402–409). IEEE Comput. Soc. doi:10.1109/ICDE.1990.113493

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

Hume, W. F. (1911). The Effects of Secular Oscillation in Egypt during the Cretaceous and Eocene Periods. *Quarterly Journal of the Geological Society*, *67*(1-4), 118–148. doi:10.1144/GSL.JGS.1911.067.01-04.05

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

Karl, R. T., & Young, J. P. (1985). Recent heavy precipitation in the vicinity of the Great Salt Lake: just how unusual. *Journal of Climate and Applied Meteorology*, *25*, 353–363.

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

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

Morgan, T. H. (1905). The relation between normal and abnormal development of the embryo of the frog. *Archiv Für Entwicklungsmechanik Der Organismen*, *19*(4), 581–587. doi:10.1007/BF02321200

Pennell, C., & Reichler, T. (2011). On the Effective Number of Climate Models. *Journal of Climate*, *24*(9), 2358–2367. doi:10.1175/2010JCLI3814.1

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

Russa, C. La, & Andrews, K. (2010). AC 2010-1093 : Managing A Digitization Project : Issues for State Agency Publications with Folded Maps. *American Society for Engineering Education*, 1–6.

Vuorinen, V., Larmi, M., Schlatter, P., Fuchs, L., & Boersma, B. J. (2012). A low-dissipative, scale-selective discretization scheme for the Navier–Stokes equations. *Computers & Fluids*, *70*, 195–205. doi:10.1016/j.compfluid.2012.09.022

Williams, K. J., Ahijevych D.A., Kessinger C.J., Saxen T.R., Steiner M., & Dettling S. (2008). A Machine Learning Approach to Finding Weather Regimes and Skillful Predictor Combinations for Short-Term Storm Forecasting. *National Center for Atmospheric Research*, 1–6. Retrieved from http://opensky.library.ucar.edu/collections/OSGC-000-000-003-270

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

Acknowledgement

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 for financial support.

Appendix A

The UAC algorithm (Hooshsadat & Za, 2012).

Algorithm 1 UAC Rule Filtering: Stage 1

```
Q = ;;U = ;;A = ;
for all i 2 Dataset do
i:ucRule = firstCorrect(i)
i:cApplic = _(i:ucRule; i)
i:uwRule = firstWrong(i)
i:wApplic = _(i:uwRule; i)
U:add(ucRule)
ucRule:covered[i:class] + +
 if (ucRule _[i] uwRule) and ucRule _ uwRule then
Q:add(ucRule)
flag(ucRule)
else
A:add(< i:id; i:class; ucRule; uwRule>)
end if
end for
```

Algorithm 2 UAC Rule Filtering: Stage 2

```
RepDAG = ;
for all < i:id; y; ucRule; uwRule>2 A do
if flagged(uwRule) then
ucRule:covered[y] □□
uwRule:covered[y] + +
else
wSet = allCoverRules(U; i:id; ucRule)
if !RepDAG:contains(ucRule) then
RepDAG:add(ucRule)
end if
for all w 2 wSet do
w:replace:add(<ucRule; i:id; y >)
w:covered + +
ucRule:incom + +
if !w 2 RepDAG then
RepDAG:add(w)
end if
end for
Q = Q:add(wSet)
end if
end for
S   set of all nodes with no incoming edges
while S 6= ; do
r = S:next() fnext removes a rule from the setg
for all <ucRule; id; y >2 r:replace do
```

```
if (r:covered[r:class] > 0) then
if id is covered then
r:covered[y] □□
else
ucRule:covered[y] □□
Mark id as covered.
end if
end if
ucRule:incom□□
if ucRule:incom = 0 then
S:add(ucRule)
end if
end for
end while
```

Algorithm 3 UAC Rule Filtering: Stage 3

```
C = ;
for all r 2 Q do
if r:covered[r:class] > 0 then
finalSet:add(r)
ruleErrors+ = computeError(r)
defClass = addDefaultClass()
defErrors = computeDefErr(defClass)
defAcc = addDefAcc(uncovered(D) □defErrors)
totalError = defErrors + ruleErrors
C:add(r; totalError; defClass; defAcc)
end if
end for
Break C from the rule with minimum error
C contains the _nal set of rules
default = defClass:get(C:size)
defApplic = defAcc:get(C:size)
                jTj
```