Flexible Workflow Description with Fuzzy Conditions

C. Zirpins†, K.Schütt† and G.Piccinelli‡

† Dept. of Computer Science(VSIS), University of Hamburg, Germany ‡ Hewlett -Packard Laboratories Bristol, UK

Abstract: To improve communication efficiency companies nowadays start making their business services available electronically over the Internet. Examples can be found in the supply-chain management, finance relations and other sectors mainly for Business-to-Business (B2B) interaction. The communication of electronic services follows specific interaction patterns that can be described through processes. A process description may be created manually or can be derived through process mining techniques from execution logs. The focus of this work is the automatic generation of branching conditions from statistical evaluations during the process mining. Our approach is to apply fuzzy logic [Zim91, Cox92, SK92] instead of crisp conditions to express uncertainty. The anticipated branching conditions are based on selected state attributes. The results are fuzzy rules suitable for a flexible description of branching conditions that allow for dynamic branching predictions.

1. Introduction.

Boolean expressions and crisp conditions are not well suited for the expression of uncertainty. The main issue addressed in this work is to derive and to describe branching conditions that enable uncertain predictions rather than deterministic flow control. The basic assumption in our work is that the process structure is known but the conditions for alternative branches are undefined. We also assume that a set of execution logs is available. These logs should provide system and process state information prior to each branching point. The internal branching decisions are likely to be based on that information. Our approach is to apply fuzzy logic [Zim91, Cox92, SK92] instead of crisp conditions to express the uncertainty. Initially we select system and process state attributes that might be relevant for the different branching choices. The anticipated branching conditions are then based on these attributes. Consequently appropriate membership functions, required in fuzzy logic terms, are derived. Finally rules for the branching conditions based on fuzzy logic are put together.

The following sections are structured according to this process. Section 2 explains why crisp conditions are not appropriate for the specific needs in our context. Section 3 provides a strategy to select relevant state attributes. Section 4 describes a way to choose the fuzzy sets that are needed for the fuzzy rule creation, which is described in section 5. In section 6 we review the overall technique and present our conclusions.

2. The need for uncertainty in Workflow Conditions

In the previous section the issue of uncertain branching conditions in connection with process mining was introduced. In this section the problem is considered in more detail and we point out the requirements for possible solutions. As an example we look at a simple alternative branching to interaction B or C after interaction A. We assume that three state attributes (a1, a2 and a3) are observable. The ranges of the attribute values in case of the two alternative branches can be overlapping, that is, the values are not completely distinct for either branch. The branching-relevant attributes indicate the branching choice only with a certain probability and do not allow for a clear decision. Further more ambiguity may occur especially within the region of overlapping ranges. I.e. two attributes indicate different branching choices, which must not invalidate the derived condition. Therefore flexible rules are needed in this case rather than Boolean expressions and crisp conditions. These rules should support a decision or allow predictions rather than enabling a deterministic process execution.

3. Finding branching-relevant attributes

In our approach we assume that the density function for sate attributes is a normal distribution. This is accompanied by the assumption that the values of one attribute lie in one single interval for each alternative branch. I.e. if the value range would actually consist of many separate intervals, the range would be considered as a single interval without separation. Many attributes, like price offers or order volumes, are likely to have such a normal distribution. Anyway, assuming a certain type of distribution is a simplification and some distributions are not well captured. These are mainly distributions with more than one maximum and multiple broken value ranges for alternative branches. Such distributions occur if, for example, one branch is chosen when a value exceeds a certain high value or falls below a smaller value, and if another branch is chosen in the middle case. For that case an approximation as a normal distribution and ranges in the form of intervals for both branching choices would not provide enough distinction for the two distributions. For the reason that the simplification is extremely helpful and most common situations are well covered, we will continue assuming the normal distribution, being aware that some branching relevant attributes are not recognised by that. This does not introduce mistakes into the derived branching conditions but it possibly reduces the capabil ity.

The normal distribution is specified by the mean and the variance, which can be estimated through the average value and the root-mean-square deviation of the observation. The range, simplified as an interval, is specified by a minimum and a maximum va lue. The interval can be estimated by simply taking the minimum and the maximum value of the observation but erroneous outliers may falsify that kind of interval estimation significantly. Therefore simple correction mechanisms like ignoring a little percen tage (e.g. one percent) of the biggest and the smallest values and calculating the interval from the remaining can bear better results.

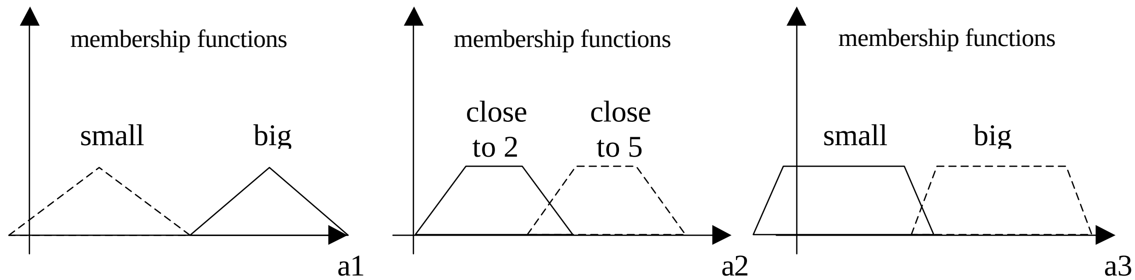

Figure 1 Fuzzy Sets for the labels of the control variables a1, a2, a3

4. Choosing fuzzy sets for branching-relevant attributes

For controlling tasks with uncertain decisions and possible ambiguity the application of fuzzy logic has proven to be appropriate. Fuzzy logic uses fuzzy sets that can be defined through membership functions. The membership function can provide a mathematical description of labels like 'big', 'small' or 'medium' concerning an attribute. The degree of membership for a certain value can vary between 0% and 100%. Fuzzy rules can then include those labels instead of crisp exp ressions. Figure 1 shows fuzzy sets for the three attributes a1, a2, a3 and related labels. The attributes become control variables in fuzzy rules like '`IF a1 is big AND a2 is close to 5 AND a3 is small THEN B follows A`'. The labels 'big', 'small' and 'close to 5' replace crisp expressions like '`a1>=10', 'a2 in(4.5; 5.5)' and 'a3<0.5'`.

Membership functions in general can be defined quite arbitrary. However in combination with fuzzy rules for controlling and decision-making fuzzy sets usually have simple membership functions with trapezoid shape. Three parameters, the top width, the bottom width and the centre specify those membership functions. Such membership functions are characterized by an interval where the membership is 100%, e dges with a linear decrease on membership and a remaining area on both sides without membership. Examples are shown in Figure 1.

From a global perspective such trapezoid membership functions suit our desire to express the confiden ce for a branching prediction, which depends on different attribute values, very well. Again we assume observations of one attribute for two alternative branches with distinct normal distributions and small interval overlaps, as illustrated in Figure 2. It is likely that there is a value range for each branch and each attribute, where the confidence is very high (say 100%) that the branch is chosen if the attribute value is in that range. Further more, in the area of the overlap, th e confidence is likely to decrease with growing distance from the 100% confidence area. Outside no indication for the branching decision exists at all, so that the confidence is zero.

The exact strategy we propose is described in the following two parts se parate for the bottom intervals and the top intervals of the membership functions. The methods we propose imitate the intuitive approach with mathematical terms.

Bottom interval

The bottom interval represents the attribute value area where at least a mini mum indication for the related branching choice is given. For example we consider the case of a binomial branching choice, where two intervals and so with two fuzzy sets for one attribute are involved. In the following we refer to the intervals that enclos e 90 or 95 percent of the observations, which are nearest to the mean, as the '90% -interval' or the '95%-interval'. In case of an overlap we might want to reduce the bottom interval in the overlapping area if one branch has much more statistical support th an the other. Apart from any overlap the minimum extent should be the width of the

95%-interval. In the 95%-interval no overlap should reduce the branch confidence to zero because clear evidence exists that the branch may occur for those values. The maximum additional extent should be until to the beginning of the 90%-range of the competing distribution, because starting from there the support for the other branch becomes very strong and if it is outside of the 95%-range of the considered branch, the own support is already very low. If the overlap is outside of the 95%-intervals, the bottom intervals are not reduced and span over the entire interval. In case of no overlap we will extend the intervals to fill the gap between them. It seems natural to give at least a little indication towards the branch whose observations are on average nearer to a newly observed value. So we extend the bottom intervals of both sets towards the middle of the gap. With those constrains the bottom interval of the fuzzy set is well defined and only the areas outside of both intervals stay unspecified. To reduce this unspecified area, we extend the outer limits of the bottom intervals to some high value, for example by adding ten times the interval size.

Top interval

The top intervals specify the area where the confidence for one branching decision is 100%. Therefore the top interval should definitely not extend the observation interval and also it should not reach significantly into an overlap area with noticeable confidence for the competitive branch. A clear minimum constraint is the distribution mean. At least in the mean a 100% confidence should always be given. The pre-selection of relevant attributes guarantees that the means of the two distributions would always have a minimum distance of one or two variance measures. So it is guaranteed that the minimum top intervals do not overlap.

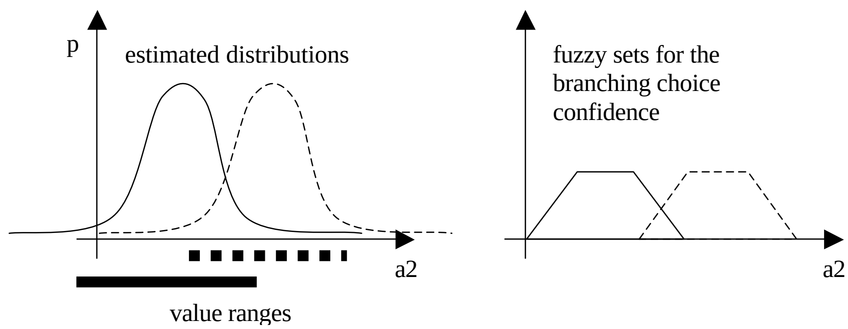

Figure 2 Attribute distributions for branches B (solid) and C (dotted)

We consider the case of a binomial branching choice in more detail. If the two observation intervals for one attribute do not overlap, the branching decision could have been made independently from other attributes. So we chose the entire observation intervals themselves as top intervals. In case of an overlap that does not include either of the distribution means, the overlapping area is excluded from the top-intervals. This fulfils the requirement not to extent the top interval into an overlap area with noticeable confidence for the competitive branch. If the overlap includes either or both of the means, then the top interval spans until to the overlapped mean even though it then reaches into the overlapping area. The reason for that inconsistency is that the statistical support for one branch at the mean point of the observations is significantly higher than for the other branch. If that would not be the case the attribute would fail in the pre-selection and would not be considered further. In any case of a one-sided interval overlap, the top interval is equal to the observation interval on the side without overlap.

The resulting fuzzy sets are supposed to be used mainly for internal branching predictions, so that no meaningful labels need to be created. The next section will explain how the fuzzy sets can be composed into fuzzy rules that replace the branching conditions.

5. Creating Fuzzy Rules for Workflow Conditions

The previous section introduced fuzzy logic and described how suitable fuzzy sets for the different attributes and branches can be derived from the statistical observations. This section shows how fuzzy rules, that replace crisp branching conditions, are composed from those sets. The fuzzy rule for one branch is composed of the fuzzy sets relating to the attribute distributions relating to this branch. Single attribute conditions in the form '`[Attribute] is member of [Fuzzy Set]`' are connected with AND to build the rule for the branch. The result is one rule for each branch, for example '`IF a1 is big AND a2 is close to 5 AND a3 is small THEN B follows A`'. The labels represent the previously derived fuzzy sets. Each rule lets us calculate a probability for one branch of an upcoming branching point based on the state attributes. The evaluation of the fuzzy rules was described in the previous s ection.

OR-concatenations are not used in the generated rules. They would be reasonable if the rules were to be designed by business experts. But OR-connected attributes do not influence the result independently and are therewith difficult to identify. We consider the branching-relevant attributes as independent and therefore derive only AND-concatenations in our approach.

To justify the use of AND-concatenations we look at a reference case. The rule we consider is `'IF a1 is big AND a2 is small THEN B follows A'`. Internally the labels 'big' and 'small' exist only as membership functions representing the probabilistic support for a branching choice. The AND-concatenation causes that the smaller membership value of a1 and a2 is returned as the confidence that B follows A. This is, in case of a1 not being 'big' or a2 not being 'small', the confidence that B follows is zero. Assuming that the membership functions for 'big' and 'small' are derived according to the previous section, this behaviour is desired, because a membership function is zero only if no statistical support for the related branch exists. As we prefer a pessimistic approach it is reasonable to return a zero confidence for one rule if at least one of the memberships is zero. This is the case only if AND-concatenations are used. In case of OR, the maximum membership would be returned, which requires all memberships to be zero to get zero as the result. The AND-concatenation also ensures that the effects of the simplified assumptions that we made con cerning the independency of branching-relevant attributes get compensated partly.

In this section we completed our methodology for the composition of fuzzy rules, which provide a flexible way to describe uncertain branching conditions. In the last section we review the overall technique and present our conclusions.

6. Conclusions.

We introduced process mining as a technique to derive process descriptions from execution logs in the context of conversation processes for electronic business services. The focus was on the automatic generation of branching conditions from statistical evaluations. It was shown that crisp conditions are not always sufficiently flexible to support probabilistic predictions and uncertain decisions. Nevertheless this flexibility was n ecessary if processes are mined from observations and not build on complete information and clear constrains. Our approach was to apply fuzzy logic instead of crisp conditions to express branching conditions with uncertainty.

References

[AGL98] Rakesh Agrawal, Dimitrios Gunaopulos and Frank Leymann: **Mining Process Models from Workflow Logs**. In Proc. Of the 6. International Conference on Extending Database Technology (EDBT), 1998

[BBB01] A. Banerji, C. Bartolini, D. Beringer, K. Govindarajan, A. Karp, H. Ku no, M. Lemon, G. Pogossiants, S. Sharma, S. Williams: **Web-Services Conversation Language (WSCL)**, Hewlett-Packard Web-Services Organization, 5/2001

[Cox92] E. Cox: **Fuzzy fundamentals**, IEEE Spectrum, vol. 29, pp. 58-61, October 1992.

[CW00] J.E. Cook and A.L. Wolf: **Discovering Models of Software Processes from Event -Based Data**, ACM Transactions on SoftwareEngineering and Methodology, 7(3):215 -249, 1998. International Journal of Intelligent Systems in Accounting, Finance and Management, 9:67 -92, 2000.

[ebX02] Web-Seite: **ebXML**, http://www.ebxml.org/, 7/2002

[Herb00] Joachim Herbst: **A Machine Learning Approach to Workflow Management**, In: R. López de Mántaras, E. Plaza (Eds.): ECML 2000, LNAI 1810, pp 183 -194, 2000

[Ros02] Web-Seite: **RosettaNet**, http://www.rosettanet.org, 7/2002

[Sch01] Guido Schimm: **Process Mining elektronischer Geschäftsprozesse**, Elektronische Geschäftsprozesse EBP 2001, Klagenfurt, Sep. 2001

[SK92] D. G. Schwartz and G. J. Klir: **Fuzzy logic flowers in Japan**, IEEE Spectrum, vol. 29, pp. 32-35, July 1992.

[WA01] A.J.M.M. Weijters and W.M.P. van der Aalst: **Rediscovering Workflow Models from Event -Based Data**. In: Hoste, V., and De Pauw, G. (Eds.), Proceedings of the Eleventh Belgian -Dutch Conference on Machine Learning (BENELEARN 2001). CNTS -Language Technology group, University of Antwerpen, pp 93 -100, 2001

[Zim91] H.-J. Zimmermann: **Fuzzy Set Theory and Its Application**, second edition, Kulwer Academic Publishers, 1991