Note and suggestion:

Overall

* I recommend using an Oxford comma.
* ‘However’, ‘Therefore’, ‘Thus’ are conjunctive adverbs. If liking two complete sentences together, a ‘semicolon’ rather than a ‘comma’ should precede conjunctive adverbs.
* Past tense should be used when referring to other people’s work.
* Table caption should be on top.
* Number indicating \subsection or \subsubsection would be good. It will make thig more obvious and you can also easily use cross-reference to refer to these sections as well.
* Be consistent with the word using in the thesis. Ex. dataset or data set
* There should be no space between number and %
* data base 🡪 database

Data

* Perhaps you can also add the data type of MHO in the table of description of alarm data. Even though it’s kind of obvious, it is better to make the description of the data as complete as possible.
* I think it would be easier to read if you add either one of this choice in the class distribution plot
  + Label x-axis should be the name of each class
  + Add number before/after each class name in the legend

Results

* In the results chapter, you used both present tense and past tense together when describing the obtained results. Please check and decide on the tense you want to use.
* For the plot of the misclassification rate of alpha, I think added dashed line for the optimal alpha would be nice.
* When describing the confusion matrix, I think it’s better to always include the number of class along with the name. The table only presents the number of class so it is not easy to follow.
* You used ‘Apriori rules or topics’ for the section. However, the caption of figure and table, and also some texts are written with ‘topics and Apriori’ instead. It’s not a big deal but I think it’s good to have consistent of the method name for the entire thesis. This applied to the SPADE as well.
* Again, I suggest you include the number of \subsection or \subsubsection for each model. Then, it would be easier when referring to this model (use cross-reference) and make it less confusing.
* It would also be nice to have a heading for the logistic regression and the DO-probit in each method. So, the reader will know that now you will be presented with another model.

4.5.1 Apriori

4.5.1.1 Logistic regression

4.5.1.2 DO-probit

4.5.2 Apriori and topics

4.5.2.1 Logistic regression

4.5.2.2 DO-probit

* Page 12: When you write about the parameter alpha, I think you should also state that alpha only has the range from 0 to 1.
* In ‘Apriori and topics’ and ‘SPACE and topics’, it would be good if you also include the total number of coefficients in the text. Even though the plot also have a label in x-axis, it is still difficult to know the exact number.

Questions:

Abstract

* Is ‘system log file’ and ‘alarm log file’ the same thing? If so, maybe you can state that as well to make it clearer.

Introduction

* Why talk about DOLDA in the introduction? Is DO-probit an extension to DOLDA? Are there any relationship between these two method?

Data

* Still not so clear about the process of bug report/ alarm log files and also description of data (Make it difficult to understand the conclusion when you have talked about the alarm logs)
* What are 40 topics?

Method

* Partially Collapsed Gibbs sampling 🡪 Why use these two parameters to collapse this parameter? Is it because of D?
* Why do we need parallelizing? Computation cost?
* Topic indicators Z\_{1:D} 🡪 Does it mean topic indicator=topic assignments?
* Do you use ‘lift’ in the analysis? If not, I don’t think it is necessary to also mention it in the thesis.
* 183 🡪 184? Because the number of days in this range Jul-Dec, Aug-Jan, Mar-Jun are 184.
* The reason of choosing Beta distribution? Is it because of characteristic of the time elapse?
* Time dependent weight term: w\_{t} and all weights: lambda\_{ij}

lamba\_{ij} = w\_{t}?

The last occurring of the alarm has highest value? max(weight)

* A time difference weight: w\_{ir}
* What is the pattern r?
* MO and alarms? What is alarms?
* Are the variables in the data highly correlated?
* DO-probit with horseshoe prior is already implemented together in the Java?
* What does log-posterior chains tell us? If it’s stable in one value, do we use that value for something?
* Why DO-probit can’t be partition..?
* How alpha (parameter for the elastic net) is normally found? Why do you say that you have insufficient data?
* The parameter of the Beta distribution for time elapse, is it Beta(8,3) that you used as an example in the thesis?

Results

* In Resulting rules section, I’m not quite clear how you discarded the rules (#rules/#unique). It would be good to clarify things a bit more.
* Is ‘topics’ a short name of ‘Observations field topics’?
* 40 topics which were obtained by LDA. Does this results come from the previous study?
* Coefficients plot, which coefficient is the intercept? Is it no.1?
* Page 24: DO-probit using topics as covariates: What about the rest of the six folds? Is it also unstable?
* Page 25: Does it mean intercepts have small values so they are in the circle shape or intercepts will actually be represented by circle no matter how high or low values they are?
* Expected (only for logistics?)
* Table summary 4.10 and Table 4.18: Expected and Nr. of expected?
* Page 34: introduce to predict?
* Page 39: DO-probit using topics and SPADE as covariates: Does the chains also have this characteristic for the remaining six folds?

Discussion

* General remarks:
  + Fold of log-posterior 0 🡪 initial value?
  + Final model?
* Data mining:
  + Why SPACE only require a maximum of three scans? in general or only for this case?
  + SPADE in R: they do support but not for this case (given in seconds over several years)?
* Predictions:
  + Which one is global or local penalty terms? horseshoe=local?
* Evaluation
  + Rule mining procedure = parameter for the Beta distribution?

Conclusion

* So, the developer group can’t access the event in the alarm log. Normally, alarm logs show symptom?
* A set of alarms is significant to determine the developer group that will solve the bug?
* None of the classes were fitted with all the expected SPAPE covariate?