Water pipe failures can not only have a great impact on people's daily life but also cause significant waste of water which is an essential and precious resource to human beings. As a result, preventative maintenance for water pipes, particularly in urbanscale networks, is of great importance for a sustainable society. To achieve effective replacement and rehabilitation, failure prediction aims to proactively find those 'most-likely-to-fail' pipes becomes vital and has been attracting more attention from both academia and industry, especially from the civil engineering field. This paper presents an alreadydeployed industrial computational system for pipe failure prediction. As an alternative to risk matrix methods often depending on ad-hoc domain heuristics, learning based methods are adopted using the attributes with respect to physical, environmental, operational conditions and etc. Further challenge arises in practice when lacking of profile attributes. A dive into the failure records shows that the failure event sequences typically exhibit temporal clustering patterns, which motivates us to use the stochastic process to tackle the failure prediction task. Specifically, the failure sequence is formulated as a self-exciting stochastic process which is, to our best knowledge, a novel formulation for pipe failure prediction. And we show that it outperforms a baseline assuming the failure risk grows linearly with aging. Broad new problems and research points for the machine learning community are also introduced for future work.

3.2 Binary classifiers using attribute information

The system is equipped with several predictive models and also allows easy extension for integrating more new models. Cox Model [Cox, 1972] (together with a new survival analysis algorithm: Multi-Task Logistic Regression (MTLR) [Yu and Baracos, 2011]), Artificial Neural Network (ANN), Logistic Regression (LGR) and Chaid Tree are available in the system which explore the massive labeled training data. However, there are still some limitations: From the practical perspective, one problem is for some pipes, especially for those constructed before 1970, of which the associated attribute information is largely incomplete, inconsistent, or unreliable. And collecting the relevant attributes information is often costive and difficult, such as estimating the water turbidity, the rainfall, and the soil type etc.; the other subtle but worth-noting issue is one cannot guarantee the current system has identified exhaustively all factors with respect to pipe failure, and sometimes training based on incomplete covariants may be misleading. From the theoretical perspective, conventional binary classifiers like ANN, Logistic Regression or Chaid tree cannot naturally explore and model the particular property for 'censored' samples compared with the Cox model and the MTLR models [Yu and Baracos, 2011] etc. As the prediction score obtained from classifiers such as ANN, Chaid Tree is not a posterior likelihood, it makes the risk measurements from fresh and salt systems are incomparable since in many cases the system user customizes and uses different models for different systems. To address this issue, the parametric Sigmoid model [Platt, 1999; Lin et al., 2007] is used in the system to calibrate the failure likelihood on an equal footing for cross-system risk ranking.

Another advantage is knowing the likelihood naturally leads to the obtain of the failure number expectation for the next year. And it is informative for the decision maker to better determine and allocate the budget in advance.