XCS229ii Literature Review

April 12, 2021

Agenda

Literature review - Agenda

- General information

Literature review - Agenda

- General information
- Sections

Literature review - Agenda

- General information
- Sections
- Examples

Literature review - Agenda

- General information
- Sections
- Examples
- Q&A

Literature Review - General Information

Literature review – Info 1

Literature review – Info 1

- Identify and disseminate relevant research papers

Literature review – Info 1

- Identify and disseminate relevant research papers
- Due on Sunday, April 25

Literature review – Info 1

- Identify and disseminate relevant research papers
- Due on Sunday, April 25
- Worth 20 points, i.e. 20% out of the total project score

Literature review – Info 1

- Identify and disseminate relevant research papers
- Due on Sunday, April 25
- Worth 20 points, i.e. 20% out of the total project score
- 4 – 5 pages document in a free of choice format

Literature review – Info 1

- Identify and disseminate relevant research papers
- Due on Sunday, April 25
- Worth 20 points, i.e. 20% out of the total project score
- 4 – 5 pages document in a free of choice format
- Submit on Gradescope as a group

Literature review – Info 1

- Identify and disseminate relevant research papers
- Due on Sunday, April 25
- Worth 20 points, i.e. 20% out of the total project score
- 4 – 5 pages document in a free of choice format
- Submit on Gradescope as a group
- 3 extra credit points possible

Literature review – Info 2

- Considered group/team changes?

Literature review – Info 2

- Considered group/team changes?
- CFs reassigned wherever needed

Literature review – Info 2

- Considered group/team changes?
- CFs reassigned wherever needed
- Check out the past projects

Literature review – Info 2

- Considered group/team changes?
- CFs reassigned wherever needed
- Check out the past projects
- Research papers behind paywalls

Literature review – Info 2

- Considered group/team changes?
- CFs reassigned wherever needed
- Check out the past projects
- Research papers behind paywalls
- Leverage the course staff

Literature review – How to identify & read papers

Stanford ONLINE

Literature review – How to identify & read papers

- Search for papers that address the challenges found previously

Literature review – How to identify & read papers

- Search for papers that address the challenges found previously

- Divide the project in subtasks and search for papers addressing the subtasks

Literature review – How to identify & read papers

- Search for papers that address the challenges found previously
- Divide the project in subtasks and search for papers addressing the subtasks
- Search for papers that address the general purpose of your system

Literature review – How to identify & read papers

- Search for papers that address the challenges found previously
- Divide the project in subtasks and search for papers addressing the subtaskss
- Search for papers that address the general purpose of your system
- **Research Paper read methodology**: https://www.youtube.com/watch?v=733m6qBH-jI

Literature review – No. of papers per group

Literature review – No. of papers per group

Group size	Minimum number of papers
one	4

Literature review – No. of papers per group

Group size	Minimum number of papers
one	4
two	6

Literature review – No. of papers per group

Group size	Minimum number of papers
one	4
two	6
three	8

Literature review – No. of papers per group

Group size	Minimum number of papers
one	4
two	6
three	8
four	10

Literature review – Papers sources

Literature review – Papers sources

- ML conferences: ICML, NeurIPS, ICLR, KDD

Literature review – Papers sources

- ML conferences: ICML, NeurIPS, ICLR, KDD

- AI blogs: Stanford, Google, Facebook, Microsoft, Deepmind

Literature review – Papers sources

- ML conferences: ICML, NeurIPS, ICLR, KDD
- AI blogs: Stanford, Google, Facebook, Microsoft, Deepmind
- Arxiv

Literature review – Papers sources

- ML conferences: ICML, NeurIPS, ICLR, KDD
- AI blogs: Stanford, Google, Facebook, Microsoft, Deepmind
- Arxiv
- Google search

Literature Review - Sections

1. General Problem/Task Definition

1. General Problem/Task Definition

- Why did you choose the papers?

1. General Problem/Task Definition

- Why did you choose the papers?
- What are the papers trying to solve and why?

1. General Problem/Task Definition

- Why did you choose the papers?
- What are the papers trying to solve and why?
- List the papers in the Reference section of the document

2. Concise Summary

2. Concise Summary

- Describe the problem addressed in each paper

2. Concise Summary

- Describe the problem addressed in each paper
- Using your own words, briefly describe the main idea of each paper

3. Compare and Contrast

3. Compare and Contrast

- Most important section

3. Compare and Contrast

- Most important section
- Clearly point out the similarities and differences of the papers

3. Compare and Contrast

- Most important section
- Clearly point out the similarities and differences of the papers
- Explain where they agree or disagree with each other

3. Compare and Contrast

- Most important section
- Clearly point out the similarities and differences of the papers
- Explain where they agree or disagree with each other
- If the papers address different subtasks, how are they related?

3. Compare and Contrast

- Most important section
- Clearly point out the similarities and differences of the papers
- Explain where they agree or disagree with each other
- If the papers address different subtasks, how are they related?
- How do the papers apply ML Theory?

4. Future Work

4. Future Work

- Make several suggestions for how the work can be extended

4. Future Work

- Make several suggestions for how the work can be extended
- Are there open questions to answer?

4. Future Work

- Make several suggestions for how the work can be extended
- Are there open questions to answer?
- Include how the papers relate to your final project idea

5. References

5. References

- List papers sorted alphabetically by title

5. References

- List papers sorted alphabetically by title
- Full author name(s)

5. References

- List papers sorted alphabetically by title
- Full author name(s)
- Year of publication

5. References

- List papers sorted alphabetically by title
- Full author name(s)
- Year of publication
- Title

5. References

- List papers sorted alphabetically by title
- Full author name(s)
- Year of publication
- Title
- Outlet (if applicable)

5. References

- List papers sorted alphabetically by title
- Full author name(s)
- Year of publication
- Title
- Outlet (if applicable)
- Free of choice format

6. Extra Credit Slack Post

6. Extra Credit Slack Post

- Post your literature review on Slack before submission

6. Extra Credit Slack Post

- Post your literature review on Slack before submission
- Take a screenshot of the post

6. Extra Credit Slack Post

- Post your literature review on Slack before submission
- Take a screenshot of the post
- Append the screenshot in the submitted document

Literature Review - Examples

Examples

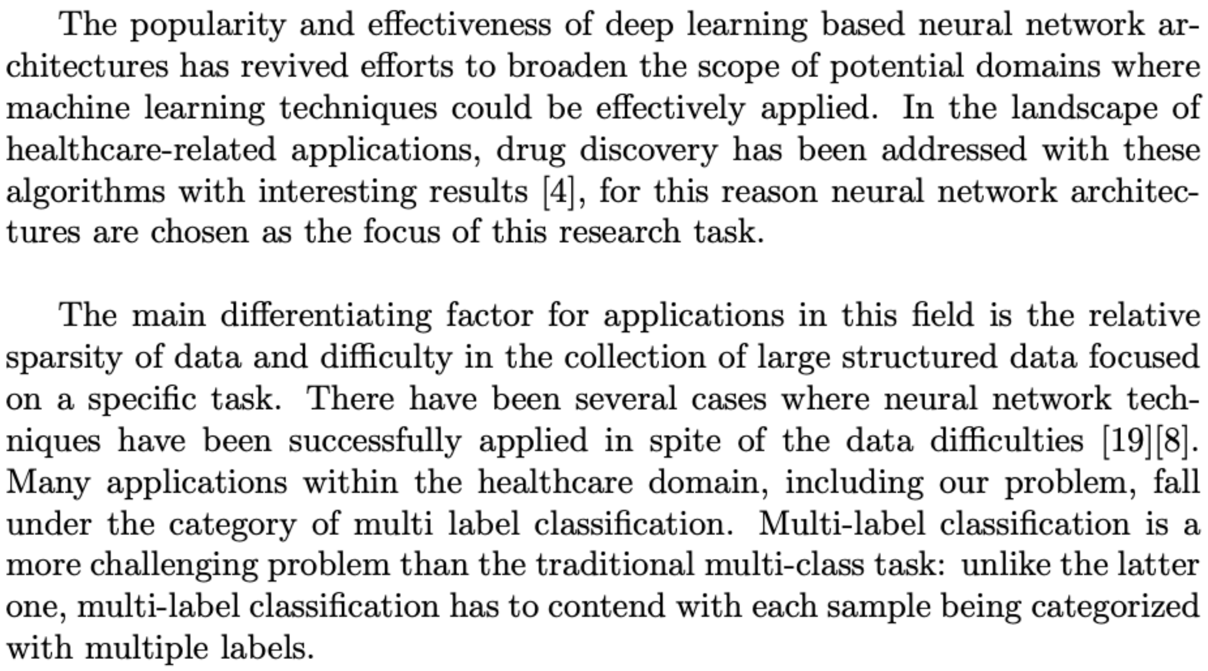
Examples

- Predicting Mechanisms of Action for Drug Discovery

Examples

- Predicting Mechanisms of Action for Drug Discovery
- OTC Derivatives Trade Reporting Advanced Analytics

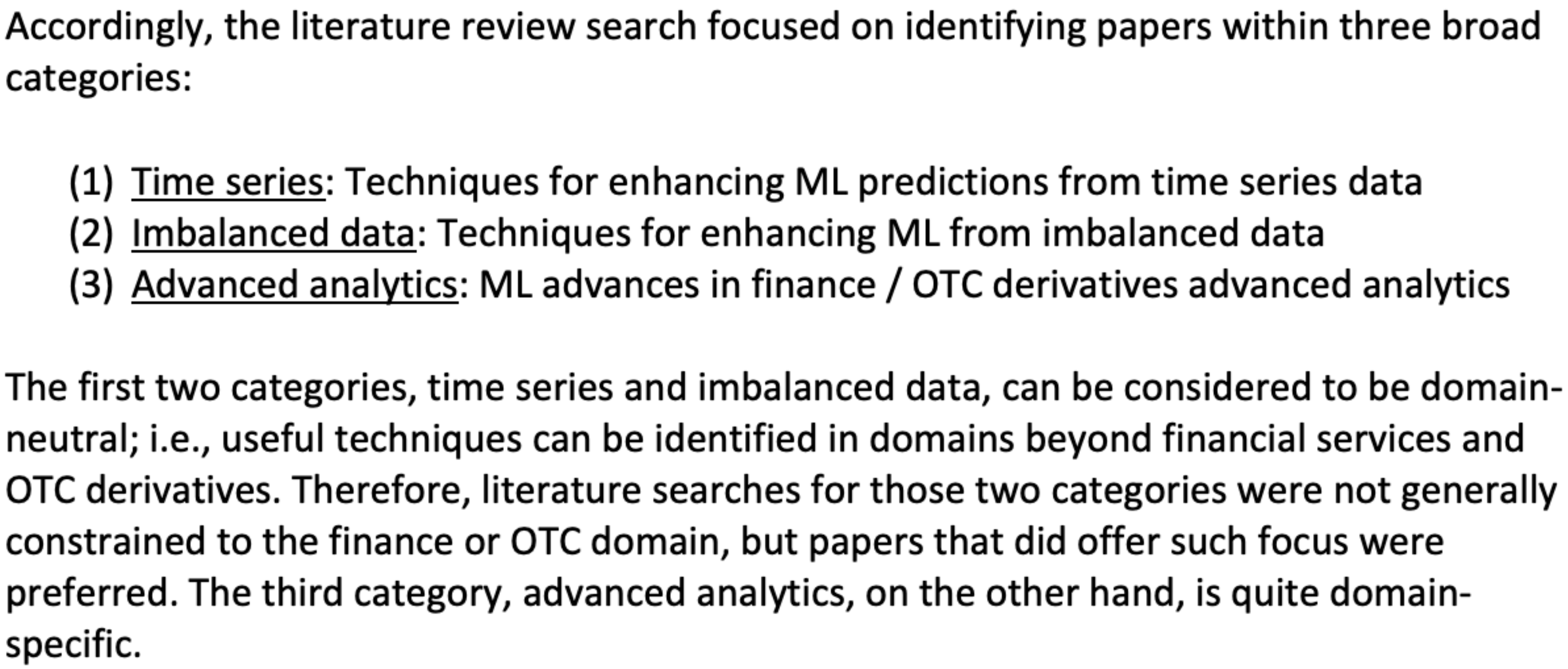
Example 1: General Problem/Task Definition



The popularity and effectiveness of deep learning based neural network architectures has revived efforts to broaden the scope of potential domains where machine learning techniques could be effectively applied. In the landscape of healthcare-related applications, drug discovery has been addressed with these algorithms with interesting results [4], for this reason neural network architectures are chosen as the focus of this research task.

The main differentiating factor for applications in this field is the relative sparsity of data and difficulty in the collection of large structured data focused on a specific task. There have been several cases where neural network techniques have been successfully applied in spite of the data difficulties [19][8]. Many applications within the healthcare domain, including our problem, fall under the category of multi label classification. Multi-label classification is a more challenging problem than the traditional multi-class task: unlike the latter one, multi-label classification has to contend with each sample being categorized with multiple labels.

Example 2: General Problem/Task Definition

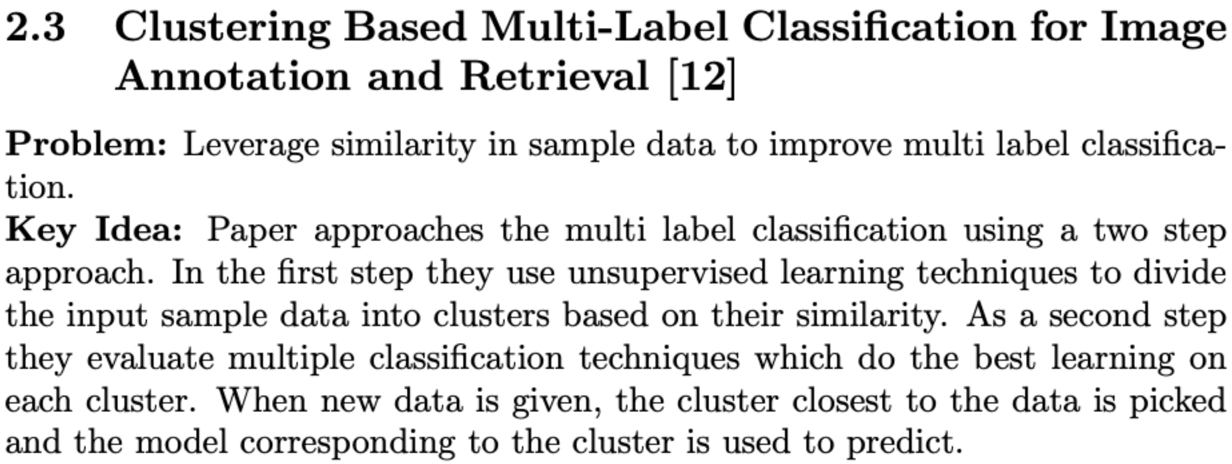


Accordingly, the literature review search focused on identifying papers within three broad categories:

(1) Time series: Techniques for enhancing ML predictions from time series data
(2) Imbalanced data: Techniques for enhancing ML from imbalanced data
(3) Advanced analytics: ML advances in finance / OTC derivatives advanced analytics

The first two categories, time series and imbalanced data, can be considered to be domain-neutral; i.e., useful techniques can be identified in domains beyond financial services and OTC derivatives. Therefore, literature searches for those two categories were not generally constrained to the finance or OTC domain, but papers that did offer such focus were preferred. The third category, advanced analytics, on the other hand, is quite domain-specific.

Example 1: Concise Summary

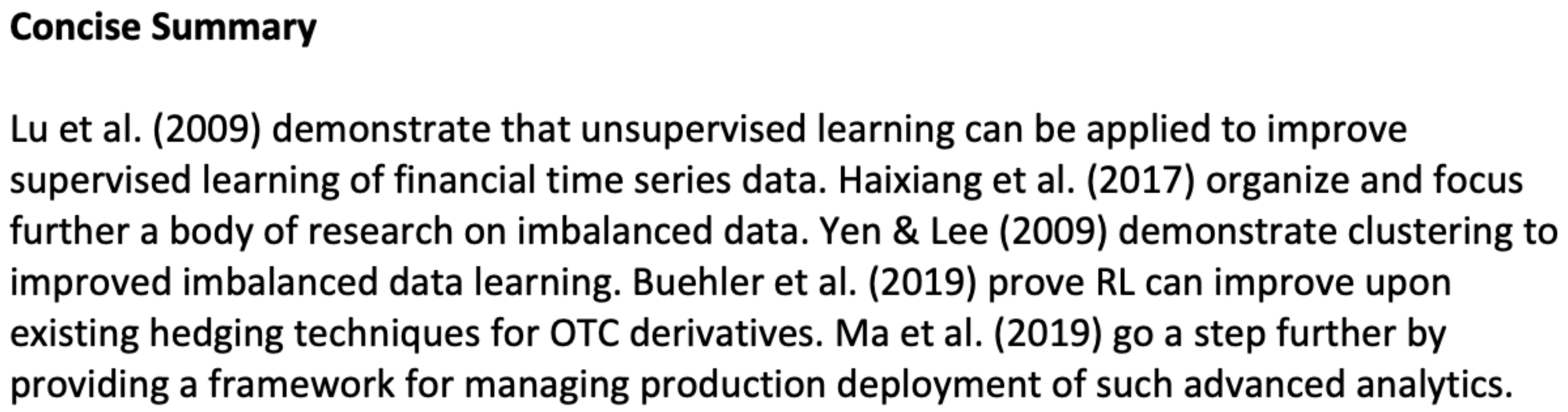


2.3 Clustering Based Multi-Label Classification for Image Annotation and Retrieval [12]

Problem: Leverage similarity in sample data to improve multi label classification.

Key Idea: Paper approaches the multi label classification using a two step approach. In the first step they use unsupervised learning techniques to divide the input sample data into clusters based on their similarity. As a second step they evaluate multiple classification techniques which do the best learning on each cluster. When new data is given, the cluster closest to the data is picked and the model corresponding to the cluster is used to predict.

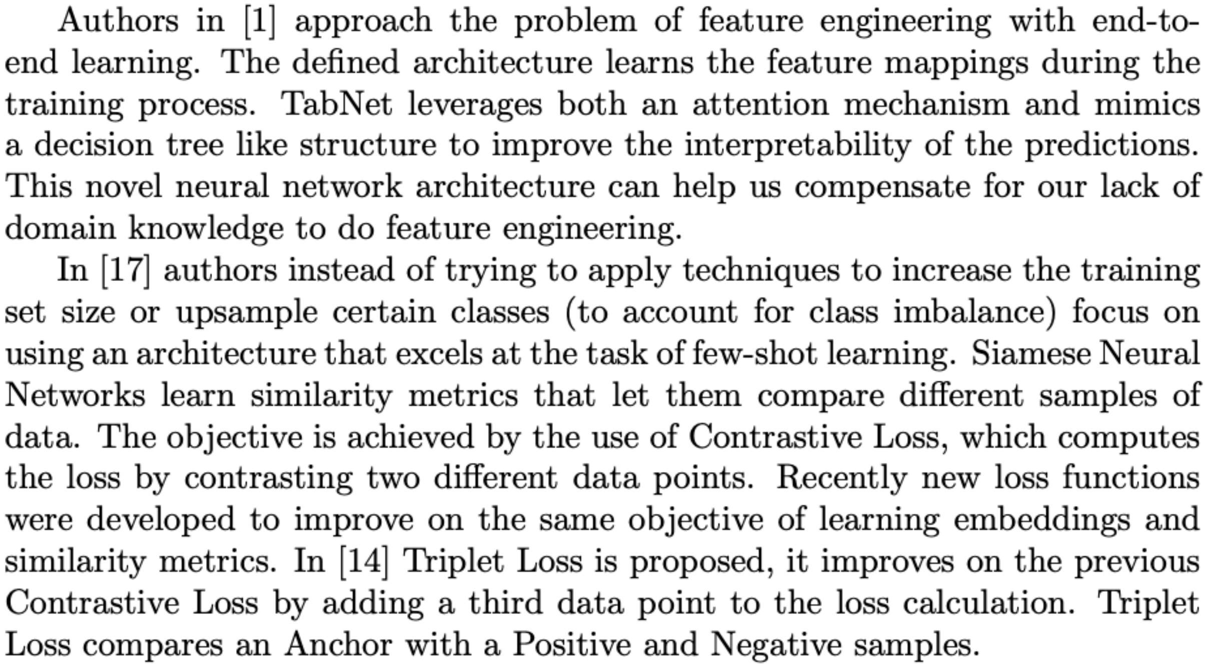
Example 2: Concise Summary



Concise Summary

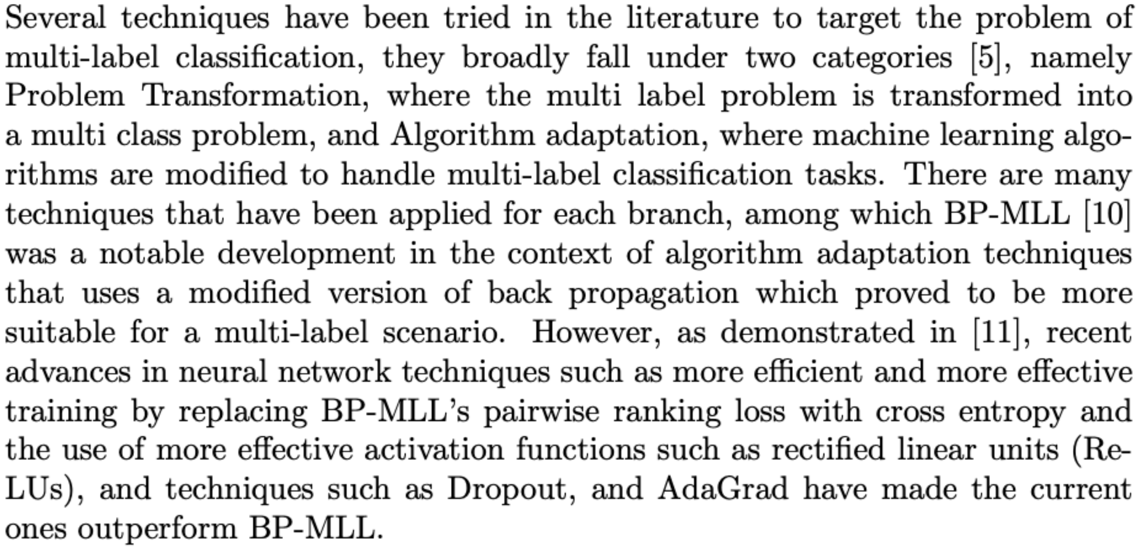
Lu et al. (2009) demonstrate that unsupervised learning can be applied to improve supervised learning of financial time series data. Haixiang et al. (2017) organize and focus further a body of research on imbalanced data. Yen & Lee (2009) demonstrate clustering to improved imbalanced data learning. Buehler et al. (2019) prove RL can improve upon existing hedging techniques for OTC derivatives. Ma et al. (2019) go a step further by providing a framework for managing production deployment of such advanced analytics.

Example 1: Compare and Contrast



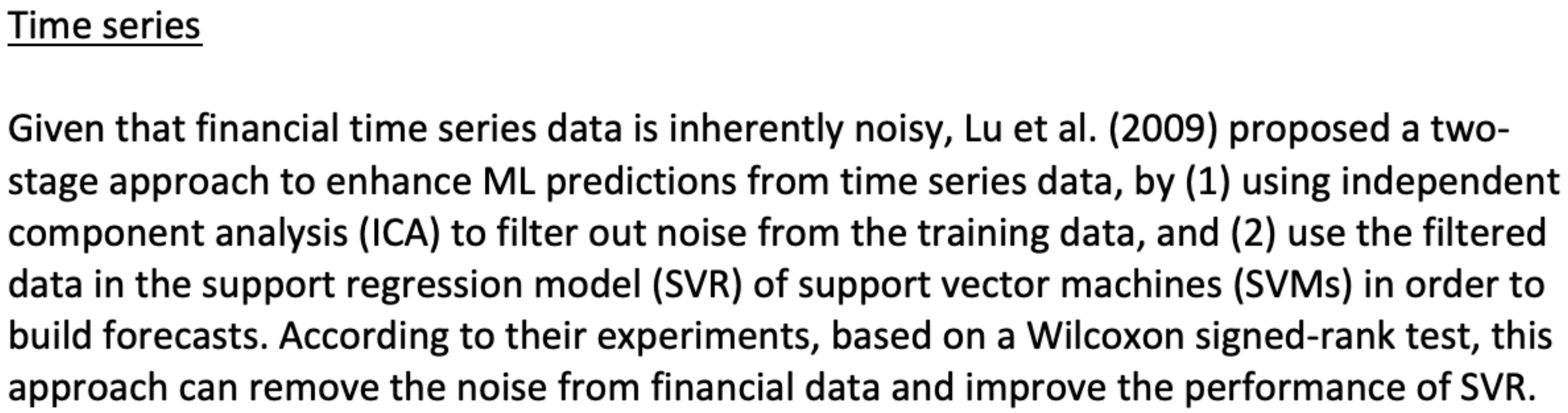
Authors in [1] approach the problem of feature engineering with end-to-end learning. The defined architecture learns the feature mappings during the training process. TabNet leverages both an attention mechanism and mimics a decision tree like structure to improve the interpretability of the predictions. This novel neural network architecture can help us compensate for our lack of domain knowledge to do feature engineering.

In [17] authors instead of trying to apply techniques to increase the training set size or upsample certain classes (to account for class imbalance) focus on using an architecture that excels at the task of few-shot learning. Siamese Neural Networks learn similarity metrics that let them compare different samples of data. The objective is achieved by the use of Contrastive Loss, which computes the loss by contrasting two different data points. Recently new loss functions were developed to improve on the same objective of learning embeddings and similarity metrics. In [14] Triplet Loss is proposed, it improves on the previous Contrastive Loss by adding a third data point to the loss calculation. Triplet Loss compares an Anchor with a Positive and Negative samples.



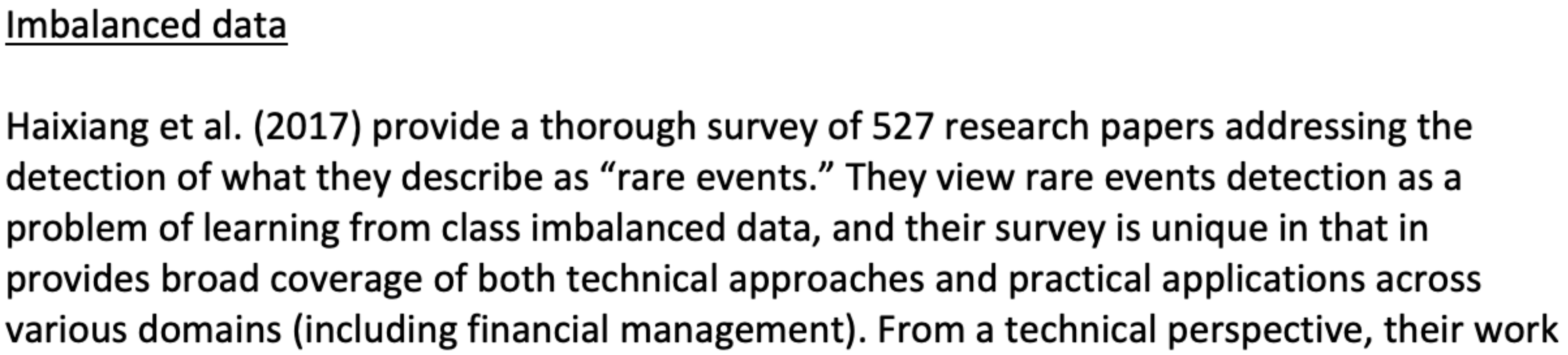
Several techniques have been tried in the literature to target the problem of multi-label classification, they broadly fall under two categories [5], namely Problem Transformation, where the multi label problem is transformed into a multi class problem, and Algorithm adaptation, where machine learning algorithms are modified to handle multi-label classification tasks. There are many techniques that have been applied for each branch, among which BP-MLL [10] was a notable development in the context of algorithm adaptation techniques that uses a modified version of back propagation which proved to be more suitable for a multi-label scenario. However, as demonstrated in [11], recent advances in neural network techniques such as more efficient and more effective training by replacing BP-MLL's pairwise ranking loss with cross entropy and the use of more effective activation functions such as rectified linear units (ReLUs), and techniques such as Dropout, and AdaGrad have made the current ones outperform BP-MLL.

Example 2: Compare and Contrast



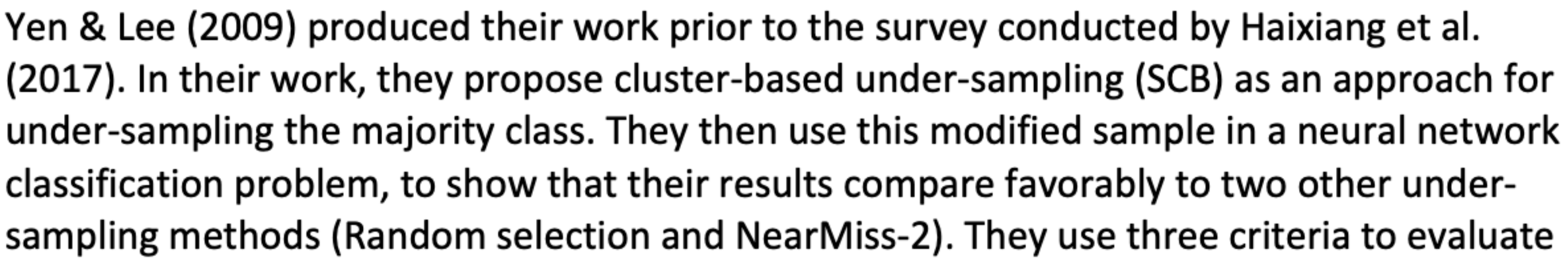
Time series

Given that financial time series data is inherently noisy, Lu et al. (2009) proposed a two-stage approach to enhance ML predictions from time series data, by (1) using independent component analysis (ICA) to filter out noise from the training data, and (2) use the filtered data in the support regression model (SVR) of support vector machines (SVMs) in order to build forecasts. According to their experiments, based on a Wilcoxon signed-rank test, this approach can remove the noise from financial data and improve the performance of SVR.



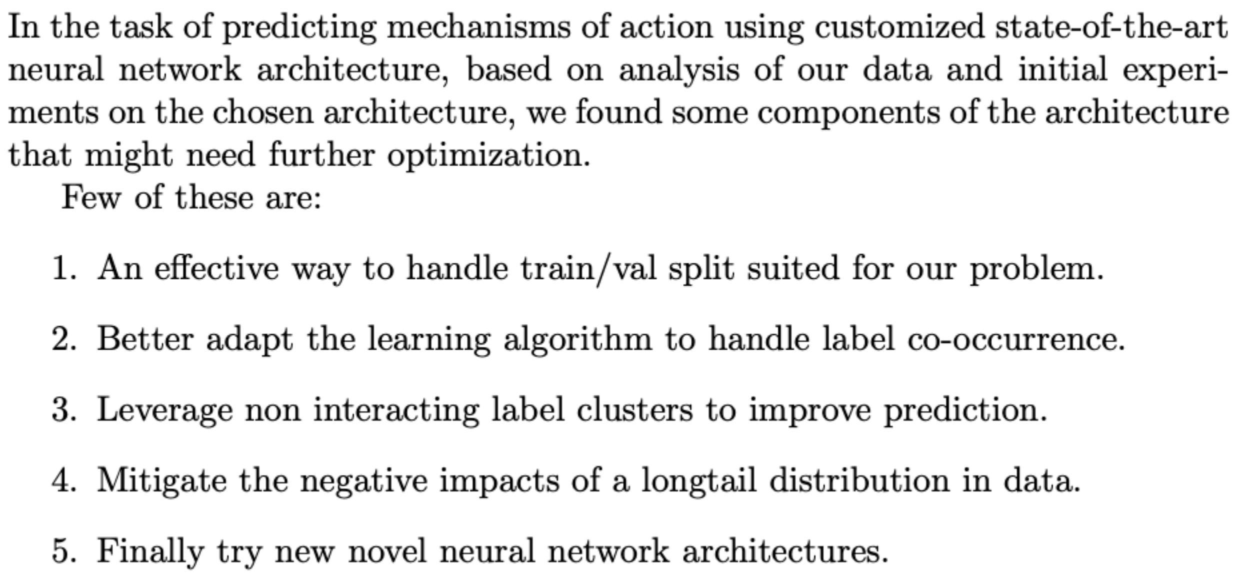
Imbalanced data

Haixiang et al. (2017) provide a thorough survey of 527 research papers addressing the detection of what they describe as “rare events.” They view rare events detection as a problem of learning from class imbalanced data, and their survey is unique in that in provides broad coverage of both technical approaches and practical applications across various domains (including financial management). From a technical perspective, their work



Yen & Lee (2009) produced their work prior to the survey conducted by Haixiang et al. (2017). In their work, they propose cluster-based under-sampling (SCB) as an approach for under-sampling the majority class. They then use this modified sample in a neural network classification problem, to show that their results compare favorably to two other under-sampling methods (Random selection and NearMiss-2). They use three criteria to evaluate

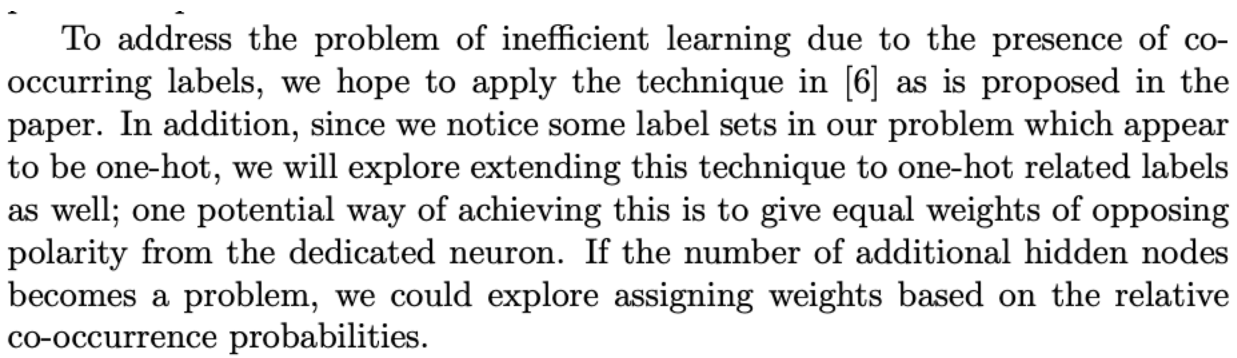
Example 1: Future Work



In the task of predicting mechanisms of action using customized state-of-the-art neural network architecture, based on analysis of our data and initial experiments on the chosen architecture, we found some components of the architecture that might need further optimization.

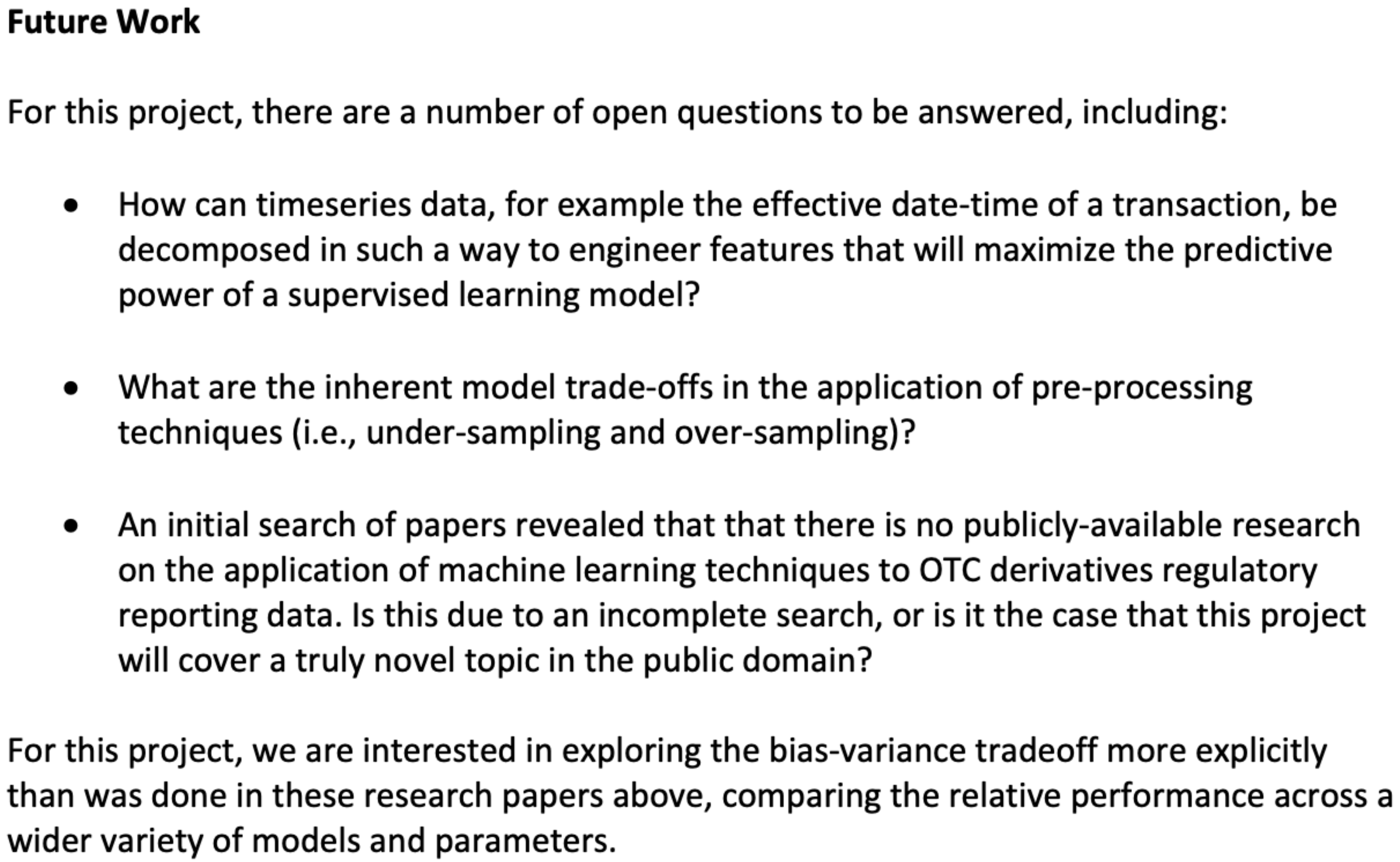
Few of these are:

1. An effective way to handle train/val split suited for our problem.
2. Better adapt the learning algorithm to handle label co-occurrence.
3. Leverage non interacting label clusters to improve prediction.
4. Mitigate the negative impacts of a longtail distribution in data.
5. Finally try new novel neural network architectures.



To address the problem of inefficient learning due to the presence of co-occurring labels, we hope to apply the technique in [6] as is proposed in the paper. In addition, since we notice some label sets in our problem which appear to be one-hot, we will explore extending this technique to one-hot related labels as well; one potential way of achieving this is to give equal weights of opposing polarity from the dedicated neuron. If the number of additional hidden nodes becomes a problem, we could explore assigning weights based on the relative co-occurrence probabilities.

Example 2: Future Work



Future Work

For this project, there are a number of open questions to be answered, including:

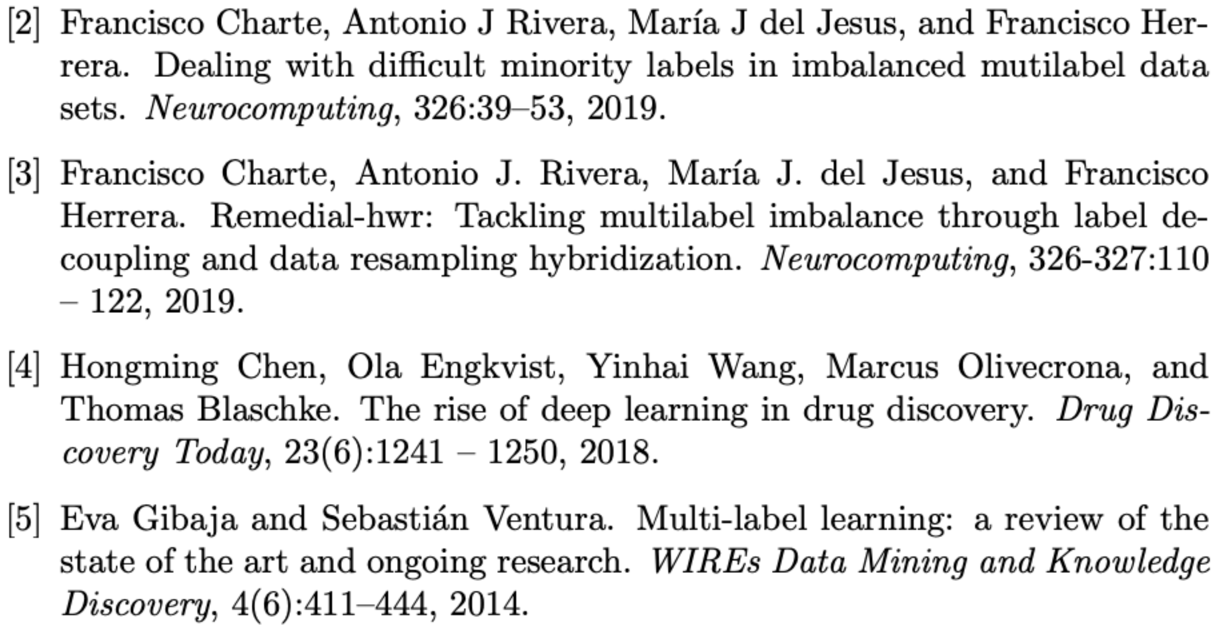
- How can timeseries data, for example the effective date-time of a transaction, be decomposed in such a way to engineer features that will maximize the predictive power of a supervised learning model?

- What are the inherent model trade-offs in the application of pre-processing techniques (i.e., under-sampling and over-sampling)?

- An initial search of papers revealed that that there is no publicly-available research on the application of machine learning techniques to OTC derivatives regulatory reporting data. Is this due to an incomplete search, or is it the case that this project will cover a truly novel topic in the public domain?

For this project, we are interested in exploring the bias-variance tradeoff more explicitly than was done in these research papers above, comparing the relative performance across a wider variety of models and parameters.

Example 1: References



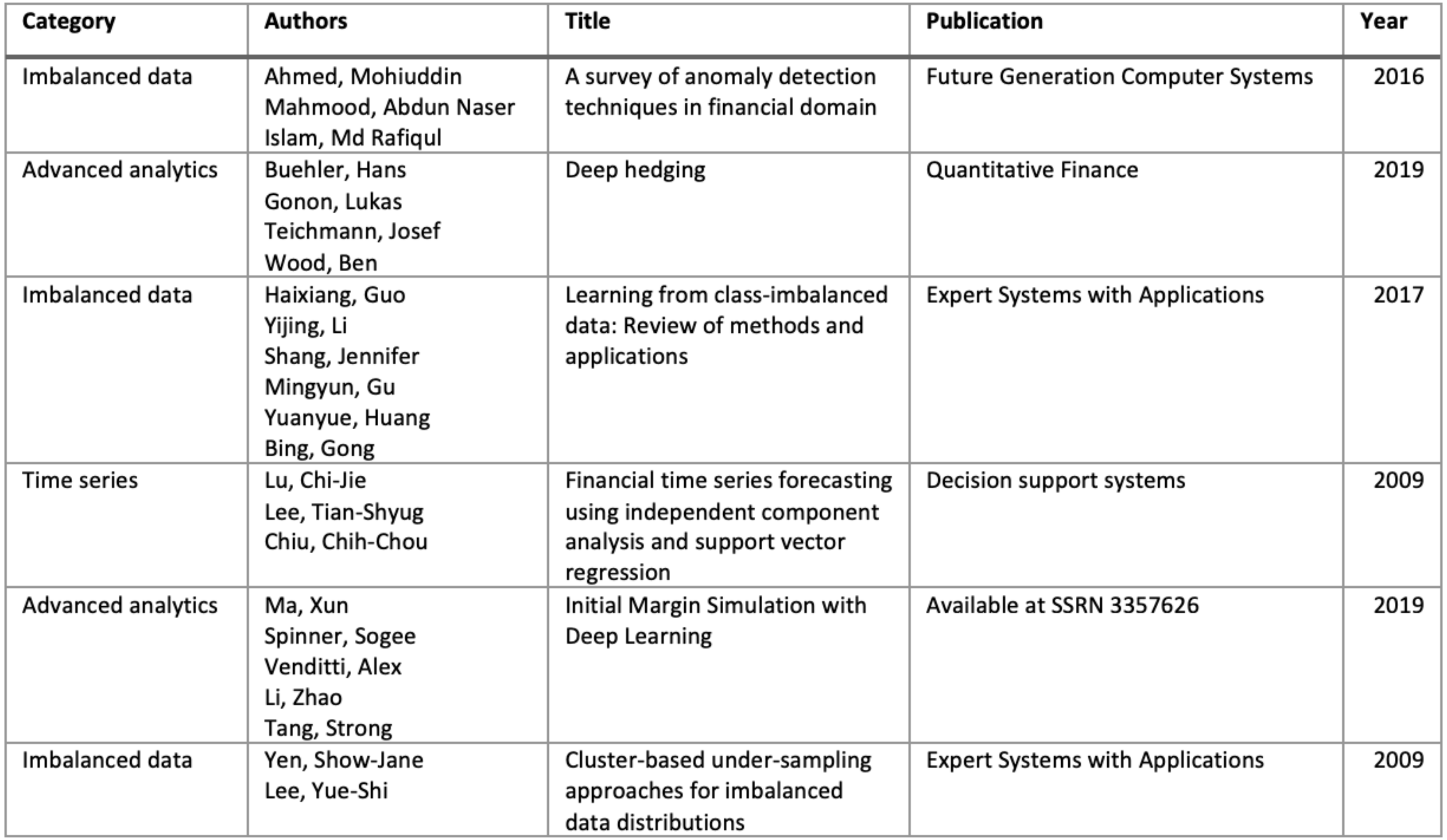
[2] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing*, 326:39–53, 2019.

[3] Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. Remedial-hwr: Tackling multilabel imbalance through label decoupling and data resampling hybridization. *Neurocomputing*, 326-327:110 – 122, 2019.

[4] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241 – 1250, 2018.

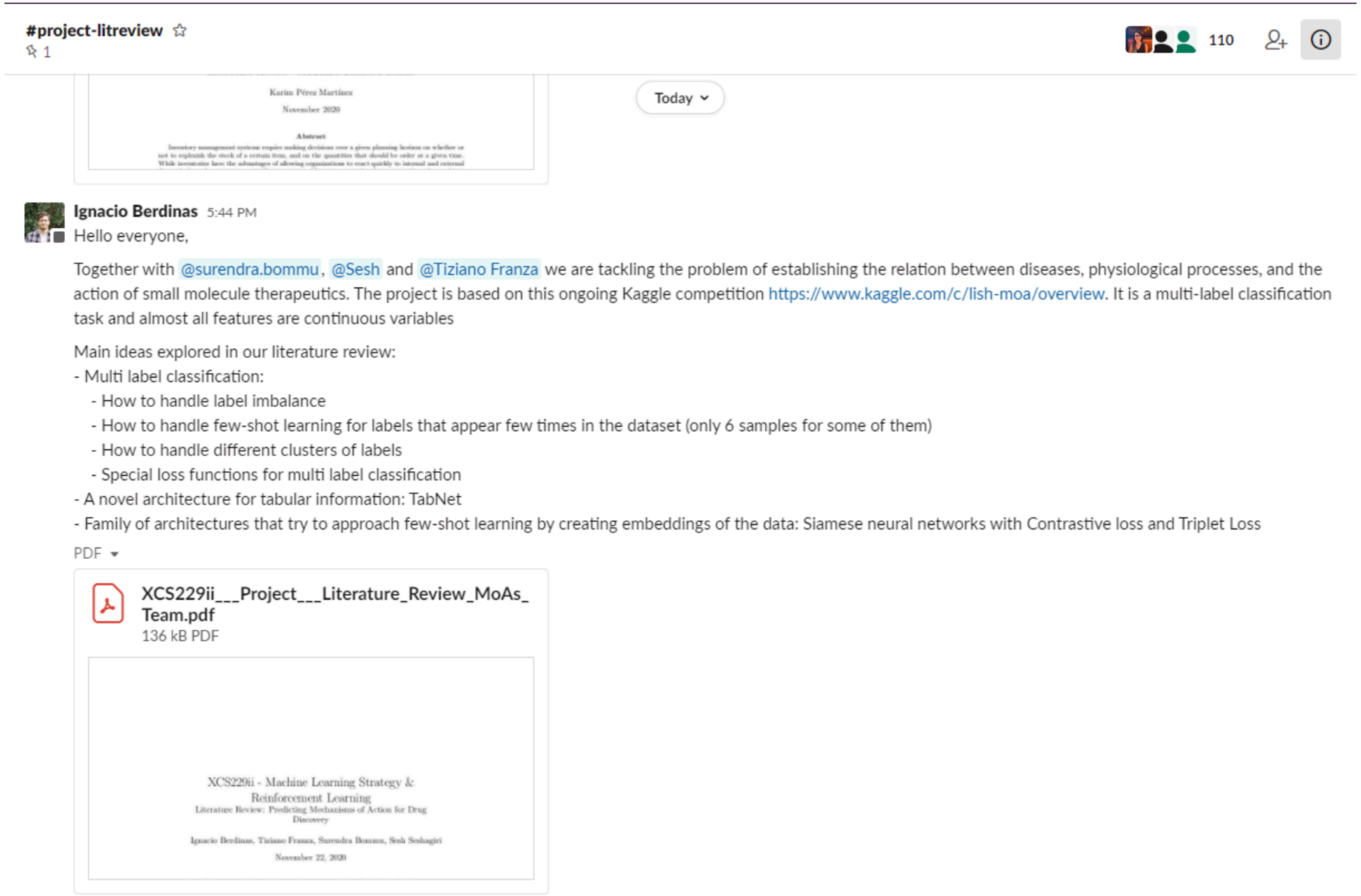
[5] Eva Gibaja and Sebastián Ventura. Multi-label learning: a review of the state of the art and ongoing research. *WIREs Data Mining and Knowledge Discovery*, 4(6):411–444, 2014.

Example 2: References



Category	Authors	Title	Publication	Year
Imbalanced data	Ahmed, Mohiuddin Mahmood, Abdun Naser Islam, Md Rafiqul	A survey of anomaly detection techniques in financial domain	Future Generation Computer Systems	2016
Advanced analytics	Buehler, Hans Gonon, Lukas Teichmann, Josef Wood, Ben	Deep hedging	Quantitative Finance	2019
Imbalanced data	Haixiang, Guo Yijing, Li Shang, Jennifer Mingyun, Gu Yuanyue, Huang Bing, Gong	Learning from class-imbalanced data: Review of methods and applications	Expert Systems with Applications	2017
Time series	Lu, Chi-Jie Lee, Tian-Shyug Chiu, Chih-Chou	Financial time series forecasting using independent component analysis and support vector regression	Decision support systems	2009
Advanced analytics	Ma, Xun Spinner, Sogee Venditti, Alex Li, Zhao Tang, Strong	Initial Margin Simulation with Deep Learning	Available at SSRN 3357626	2019
Imbalanced data	Yen, Show-Jane Lee, Yue-Shi	Cluster-based under-sampling approaches for imbalanced data distributions	Expert Systems with Applications	2009

Example 1: Extra Credit Slack Post



Karina Pérez Martínez

November 2020

Abstract

Inventory management systems require making decisions over a given planning horizon on whether or not to replenish the stock of a certain item, and on the quantities that should be order at a given time. While inventories have the advantages of allowing organizations to react quickly to internal and external

Ignacio Berdinas 5:44 PM

Hello everyone,

Together with @surendra.bommu, @Sesh and @Tiziano Franza we are tackling the problem of establishing the relation between diseases, physiological processes, and the action of small molecule therapeutics. The project is based on this ongoing Kaggle competition https://www.kaggle.com/c/lish-moa/overview. It is a multi-label classification task and almost all features are continuous variables

Main ideas explored in our literature review:
- Multi label classification:
  - How to handle label imbalance
  - How to handle few-shot learning for labels that appear few times in the dataset (only 6 samples for some of them)
  - How to handle different clusters of labels
  - Special loss functions for multi label classification
- A novel architecture for tabular information: TabNet
- Family of architectures that try to approach few-shot learning by creating embeddings of the data: Siamese neural networks with Contrastive loss and Triplet Loss

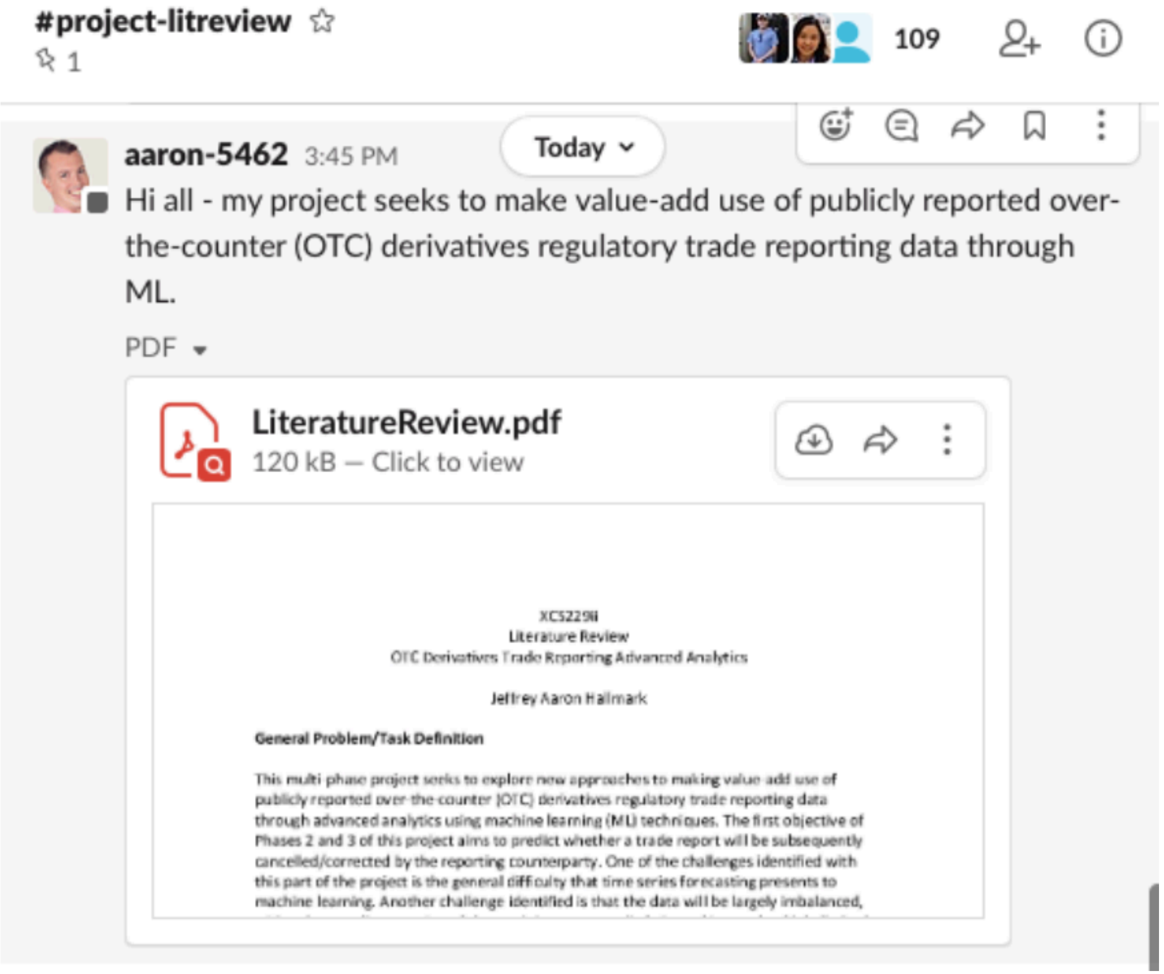
XCS229ii___Project___Literature_Review_MoAs_Team.pdf
136 kB PDF

XCS229ii - Machine Learning Strategy & Reinforcement Learning
Literature Review: Predicting Mechanisms of Action for Drug Discovery

Ignacio Berdinas, Tiziano Franza, Surendra Bommu, Sesh Seshagiri

November 22, 2020

Example 2: Extra Credit Slack Post



aaron-5462 3:45 PM

Hi all - my project seeks to make value-add use of publicly reported over-the-counter (OTC) derivatives regulatory trade reporting data through ML.

LiteratureReview.pdf
120 kB — Click to view

XCS229II
Literature Review
OTC Derivatives Trade Reporting Advanced Analytics

Jeffrey Aaron Hallmark

General Problem/Task Definition

This multi phase project seeks to explore new approaches to making value-add use of publicly reported over-the-counter (OTC) derivatives regulatory trade reporting data through advanced analytics using machine learning (ML) techniques. The first objective of Phases 2 and 3 of this project aims to predict whether a trade report will be subsequently cancelled/corrected by the reporting counterparty. One of the challenges identified with this part of the project is the general difficulty that time series forecasting presents to machine learning. Another challenge identified is that the data will be largely imbalanced,

Literature Review - Summary

Literature Review - Key takeaways

- Check Prof. Ng’s video

Literature Review - Key takeaways

- Check Prof. Ng's video
- Cover all sections and at least the minimum of papers required

Literature Review - Key takeaways

- Check Prof. Ng's video
- Cover all sections and at least the minimum of papers required
- Think about the team structure

Literature Review - Key takeaways

- Check Prof. Ng’s video
- Cover all sections and at least the minimum of papers required
- Think about the team structure
- Sharing gets rewarded

Literature Review - Key takeaways

- Check Prof. Ng’s video
- Cover all sections and at least the minimum of papers required
- Think about the team structure
- Sharing gets rewarded
- Browse through past projects

Literature Review - Key takeaways

- Check Prof. Ng’s video
- Cover all sections and at least the minimum of papers required
- Think about the team structure
- Sharing gets rewarded
- Browse through past projects
- Don’t hesitate to ask for help

Q&A

Stanford | ONLINE