**Objective**

The primary task in tax operation department is classifying the withholding tax incurred in each transaction payment. Normally, the process is to identify the service or work in those transaction based on the invoice description, purchase order and other supporting document. Then from those knowledge we identify whether it’s an object to tax liability and then determine the rate and the article of those tax liability.

This paper attempts to explore the possibility to derive the decision tree and decision rule in those classification process in hope it will open the opportunity to partially automate those process. The approach to derive the decision tree will be done using statistical learning (machine learning).

**Approach**

Statistical learning is subfield in statistic that the application objective is to extract information from data. Mainly there is two major category, supervised learning and unsupervised learning. Supervised learning is used when we have data feature or variable and its classification. Supervised learning is used to train the model using training data, which is historical data and use those model to predict the classification of future data which we haven’t yet know the classification. In summary the supervised learning is used to predict the future data based on previous data. The other type of statistical learning is the unsupervised learning is used when the data don’t have particular category or group and the learning process is aimed to group those data together based on particular characteristic in hope to gain meaningful information.

For this paper we will use supervised learning. The feature in those data is the description of transaction recorded in Oracle and the classification is the type of service and tax article applied to those transaction. In supervised learning there are many model that can be used to predict the the classification of future data but only decision tree and decision rule has the human readable reasoning process. In other model such as logistic regression, neural network, and SVM(support vector machine) the reasoning in the model is in form of coefficient in mathematical equation. So it hold no value in our purpose.

**Process**

In this paper the data we used is the one year historical data of tax payment. Ranging from September 2015 to August 2016. The reasoning is because there is a major change in tax regulation at July and August where Telkomsel was appointed as WAPU (wajib pungut) and also appointed to withheld PPh 22 for the payment of goods which previously wasn’t an object of those tax liability. So it’s wise to use the data after those implementation so it will accurately reflect the current and possible future transaction data.

From those data we will use 2 primary feature, vendor name and invoice description. The classification group is the appended combination of tax article, tariff and type of service. For example if the tax article is PPh 23, its tariff is 2% and the type of service is “sewa dan penghasilan lain sehubungan dengan harta”, then the appended classification will be “PPh 23 2% sewa dan penghasilan lain sehubungan dengan harta”.

Before goes deeper into the process, first we need to evaluate what data we have in hand and the possible processing method suitable for those data. Mainly the processed data is in form of text and in human readable language, so we will use NLP(Natural Language Processing). NLP is a method to process the human readable text and communication form, either in written text or audio language to the format processable by the computer for the statistical learning process. In this process we will use it to process the written text of invoice description and map it into the format processable by the statistical learning algorithm. This result is in form of a matrix that contains information whether those invoices contain particular word or not.

After the data is ready to be processed the algorithm will learn based on those data and evaluate which word affect most to the classification. To simply put the algorithm is split the data into two group based on whether it contain particular word or not and calculate where the entropy drop the most. Entropy here mean the variance (rate of error) in the data for those particular group. Word that drop the entropy most mean that those particular word is significant for the decision process. The process executed recursively until the variance reach bottom limit where there is no more room for improvement and we get all the keyword affected the classification process.

The result will be in the form of list of word that affect the decision process and the rate of error of those model. In complex transaction we realize that rely the classification modelling only based on invoice description won’t be enough and the result won’t be optimum.

The output rule of the learning process is the rule deducted by model based on those descriptions. This paper will only focus on the vendors that have frequent transaction. Several example of vendor actually only have one type of service such as in this list, so the model actually don’t achieve anything for these type of transaction.

|  |  |
| --- | --- |
| Vendor Name | Rules: |
| Nokia Solutions and Networks Indonesia | PPh Pasal 4 0.04 Jasa Pelaksana Konstruksi (Bukan Pengusaha Konstruksi) |
| Ericsson AB | PPh Pasal 26 0.15 Royalti-COD |
| Hariff Daya Tunggal Engineering | PPh Pasal 4 0.03 Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Wiraky Nusa Telekomunikasi | PPh Pasal 4 0.03 Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Westindo Esa Perkasa | PPh Pasal 4 0.03 Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Kinarya Alihdaya Mandiri | PPh Pasal 23 0.02 Jasa penyedia tenaga kerja |
| Profesional Telekomunikasi Indonesia | PPh Pasal 23 0.02 Sewa dan penghasilan lain sehubungan dengan penggunaan harta |

On other hand for the type of transaction that varies greatly on its nature cause the model to generate large number of rule, the highest reach up to 47 rules and have high error rate 16% to 38%. It indicate that those type of transaction need to be predicted with another kind of feature other than invoice description.

Things get interesting for the transaction have multiple tax classification but with the complexity relatively low. The model generate rule that some of them are quite reasonable from the human perspective.

|  |  |
| --- | --- |
| Vendor Name | Rules: |
| Huawei Tech Investment | Rule 1: |
| Core = Yes |
| -> class PPh Pasal 23 15% Royalti |
| Rule 2: |
| core = No |
| sis = No |
| -> class PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Rule 3: |
| sis = Yes |
| -> class PPh Pasal 4 4% Jasa Perencanaan Konstruksi |
| Default class: PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Ericsson Indonesia | Rule 1: |
| hoc150438 = Yes |
| -> class PPh Pasal 23 2% Jasa perantara dan atau keagenan |
| Rule 2: |
| hoc150438 = No |
| -> class PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Default class: PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| ZTE Indonesia | Rule 1: |
| hoc141237 = Yes |
| -> class PPh Pasal 23 2% Jasa perantara dan atau keagenan |
| Rule 2: |
| certificate = Yes |
| -> class PPh Pasal 23 2% Jasa perantara dan atau keagenan |
| Rule 3: |
| certificate = No |
| hoc141237 = No |
| -> class PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Default class: PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Abhimata Citra Abadi | Rule 1: |
| ibc = Yes |
| -> class PPh Pasal 23 2% Sewa dan penghasilan lain sehubungan dengan penggunaan harta |
| Rule 2: |
| ibc = No |
| -> class PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |
| Default class: PPh Pasal 4 3% Jasa Pelaksana Konstruksi (Menengah & Besar) |

From those result, several rule were categorized by the description of the invoice while the others are by the hoc number or the purchase order number. Categorizing by hoc number is quite reasonable because similar job tend to have the same purchase order. On the other hand, rule deduced based on description sometimes lead to the decision that don’t make sense to the human reader, such as in the case of Huawei. In the case of Huawei, a transaction is classified as a royalty if it contain keyword “core”. This result doesn’t make sense to the human reader, but if we look to those license transactions, keyword “core” always followed by “license”. So from the human perspective, we will classify a transaction as license if it contain keyword “license”. But unfortunately, alphabetical order of core come before license so it picked by the model as rule. On the other hand, the keyword “sis” is stand for “site investigation survey” and the model already correctly set it as a rule to classification.

In the case of Ericsson Indonesia, the model use hoc number as the rule. If we look into the historical data, those hoc number hold for the land certification transaction, so the result is quite reasonable to set it as a rule. Another reason as why the model use hoc number instead of keyword is because for several transactions the invoice description recorded in English while several other in bahasa. So the model considers that classifying in hoc number will be more accurate. ZTE have the similar situation as Ericsson Indonesia, but it uses the combination of keyword, and hoc number.