Development of Prediction Model for Linked Data based on the Decision Tree – for Track A, Task A1

Dongkyu Jeon and Wooju Kim

Dept. of Information and Industrial Engineering, Yonsei University, Seoul, Korea

jdkclub85@gmail.com, wkim@yonsei.ac.kr

Abstract. In this paper, we explain the detail analysis procedure of submission 1(Previous predicted results submission) of Task A1. We are trying to induce decision tree models to predict pc:numberOfTenders. Since the type of target attribute is non-negative integer value, we use the variance reduction as the attribute selection criteria. Input attributes are defined based on structure information of Public Contracts Ontology. We use the description logic constructors to properly represent a meaning of structure information of training data. Among all instances of the contract class, we make 10 different input data sets through random sampling method. The procedure of decision tree learning is performed by using SAS E-miner, and attribute selection criteria is variance reduction. Final prediction results of test data are the average of selected decision tree models except few models which have extremely low R-Square value.

Keywords: Decision Tree, Linked Data, Semantic Web, Machine Learning

1 Introduction

To predict the value of 'pc:numberOfTenders' of Task A1, classification algorithm for ordinal target attribute is required to induce the prediction model. Decision Tree algorithm [6] is one of the most popular classification method to solve a prediction problem. There are many previous researches about Decision Tree algorithms for structured data [1,4,5]. However, since these algorithms can only be learning on categorical target attribute, it is not appropriate to apply to Task A1 problem.

In this paper, firstly we generate the single table form input data based on the several attributes which are defined based on the schema of Public Contracts Ontology. After that, we induce decision tree models whose attribute selection criteria is variance reduction. With this research approach, we can induce the decision tree model for ordinal target attributes and also possible to use both nominal and interval type input attributes.

The remainder of this paper is organized as follows. In section 2, the generation procedure of input attributes is discussed. Section 3 describes the detail experiment procedure about pre-processing of input data and decision tree learning. Finally, section 4 presents conclusions and limitations of our work.

2 Input Attributes Generation

First, we need to define input attributes for decision tree learning. To reflect the structural information of Linked data, we use the concept of the refinement [1,4,5] which is used to represent the characteristic of instances from ontology by using the Description Logic constructors[2].

Input attributes for decision tree are generated based on the schema of ontology as described in Fig 1. According to both training and test data sets, we select properties and classes which are commonly appeared in both data sets. As we know, there exist much more information about contracts in training data, but it is useless when test data doesn't have matched information. Therefore only 10 properties and 6 classes are used for generating input attributes.

The list of final input attributes and its definitions are presented in Table 1. All input attributes except the target attribute are defined based on the description logic constructors. Some of attributes such as the schema:addressLocality, skos:notation are indirectly related to contract class. These attributes may have no values when the contract instance doesn't have the value of pc:location or pc:mainObject. In this case, '*none*' is filled in the missing value of attributes.

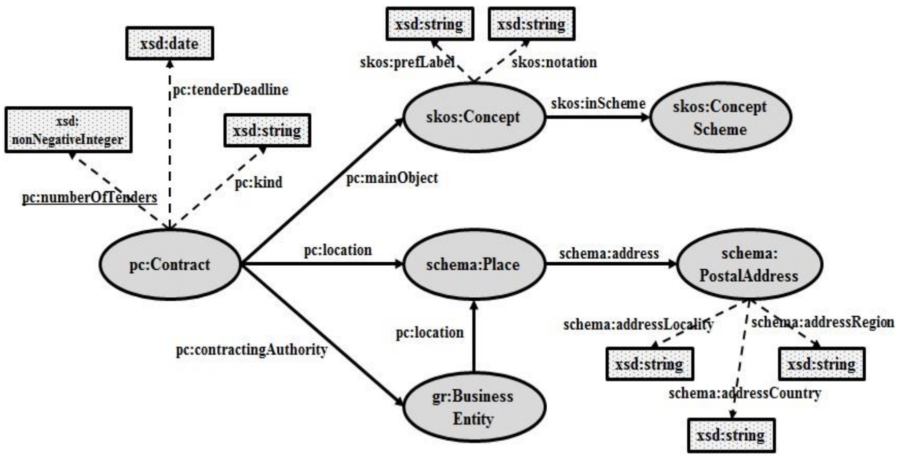


Fig. 1. Refined schema of Public Contract Ontology

3 Experiment

In this section, we present experiment procedure to learn decision tree models. This procedure is separated into two sub procedures; Firstly, we discuss about preprocessing of input data for decision tree learning. After that, decision tree learning procedure and its results are explained.

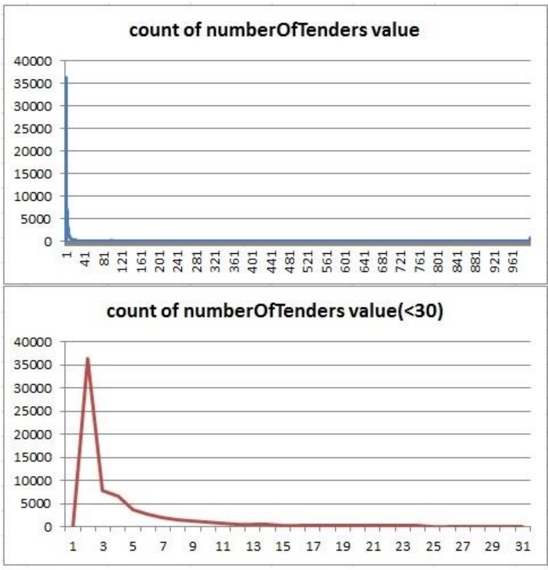
Table 1. The list of input attributes and its definition

Attribute name	Definition
pc:numberOfTenders	Number of tenders
∃ pc:contractingAuthority.(*resource*)	The contract has the *resource* as the value of pc:contractingAuthority.
∃ pc:location.TOP	Location value exists or not.
schema:addressCountry.(*value*)	The contract has a resource as the value of pc:location, and its schema:addressCountry is *value*.
schema:addressRegion.(*value*)	The contract has a resource as the value of pc:location, and its schema:addressRegion is *value*.
schema:addressLocality.(*value*)	The contract has a resource as the value of pc:location, and its schema:addressLocality is *value*.
pc:kind.(*value*)	The pc:kind value of contract is *value.*
skos:mo_notation.(*value*)	The contract has a resource as the value of pc:mainObject, and its skos:notation is *value*.
skos:mo_prefLabel.(*value*)	The contract has a resource as the value of pc:mainObject, and its skos:prefLabel is *value*.
∃ skos:mo_inSchema.(*resource*)	The contract has a resource as the value of pc:mainObject, and it has the *resource* as the value of skos:inSchema.

3.1 Preprocessing of Training Data

There are more than 70000 contract instances in training data, and each contract has a value of pc:numberOfTenders. The distribution of values is given in Fig 2. As described in details of distribution in Fig 2, 96% of contract's values of pc:numberOfTenders are less than 30. Besides almost 50% of contracts have '1' as value of pc:numberOfTenders. To reduce the effect of these dominant contracts to learning correct decision tree model, we generates ten different input data sets which are sets of randomly sampled 1500 contract instances from training data set. The sample size is determined based on the number of contracts which have the value of pc:kind (http://purl.org/procurement/public-contracts#kind) property. There are only 1683 contracts have the value of pc:kind, but it is one of the information that training and test dataset have in common. Therefore, among the ten different input data sets, five of them are randomly sampled from the set of instances which have the value of pc:kind. Other input data sets are sampled from the set of all contract instances of training data set.

All of input data generating procedure is performed by Java based application. We used Jena [3] to handle given RDF data, and inferred extra information which are not contained in original data by using the reasoner provided by Jena.



Value	count of numberOfTenders value	Proportion
0	110	96%
1	36285	
2	7829	
3	6551	
4	3602	
5	2790	
6	1982	
7	1379	
8	1165	
9	885	
10	872	
11	605	
12	503	
13	439	
14	333	
15	327	
16	245	
17	206	
18	186	
19	164	
20	200	
21	221	
22	151	
23	171	
24	94	
25	120	
26	72	
27	84	
28	77	
29	82	
30	87	
>30	2726	4%

Fig. 2. Distribution of values of pc:numberOfTenders

3.2 Decision Tree Learning

For each sampled input data, we induce decision tree by using decision tree learning module of SAS E-miner. Since the type of target attribute is ordinal, we use a variance reduction method as the splitting criterion. Input data is partitioned into 80% of training set and 20% of validation set. Results of experiments are shown in Table 2. Generally, R-Squared and ASE (Average Squared Error) value can explain the goodness of the regression decision tree.

Table 2. Results for experiment

	Training		*Validation*	
Decision Trees	*R-Squared*	*ASE*	*R-Squared*	*ASE*
Tree 1	**0.322**	**7181.47**	**0.18**	**13580.168**
Tree 2	0.085	9420.199	0.089	9034.793
Tree 3	**0.277**	**9156.532**	**0.163**	**8564.653**
Tree 4	**0.502**	**4450.398**	**0.132**	**9235.575**
Tree 5	**0.227**	**9980.726**	**0.403**	**12053.118**
Tree 6	**0.435**	**13.397**	**0.129**	**32.35**
Tree 7	**0.344**	**21.533**	**0.271**	**26.28**
Tree 8	**0.62**	**21.186**	**0.282**	**32.73**
Tree 9	0.444	18.553	0.049	22.55
Tree 10	**0.389**	**18.22**	**0.327**	**13.678**

Tree 1 ~ 5 are induced from input data sets without pc:kind attribute. *Tree 6 ~ 10* are generated on the input data with pc:kind attribute. As we can see, the scores of R-Squared value have no big difference. However, average squared error is much different in between decision trees based on the input data set with pc:kind (*Tree 1 ~ 5*) and without pc: kind(*Tree 6 ~ 10*). Fig. 3 shows the sub-tree which is condensed to 3 depth from root node of *Tree 10*. A notation value of pc:mainObject is firstly selected as the significant classifying attributes for decision tree. Information of contract about main object, contracting authority and its address are used to classify remain contract instances. There are 18 decision rules from full size *Tree 10*, and one of decision rules is described in Fig. 4. This rule can classify contracts based on its local address, contracting authority and notation of main object.

We select some decision tree models based on the score of experiment results. Decision trees in bold are finally selected models to predict the test data. The prediction value of test data is the average value of classifying results of each selected decision tree.

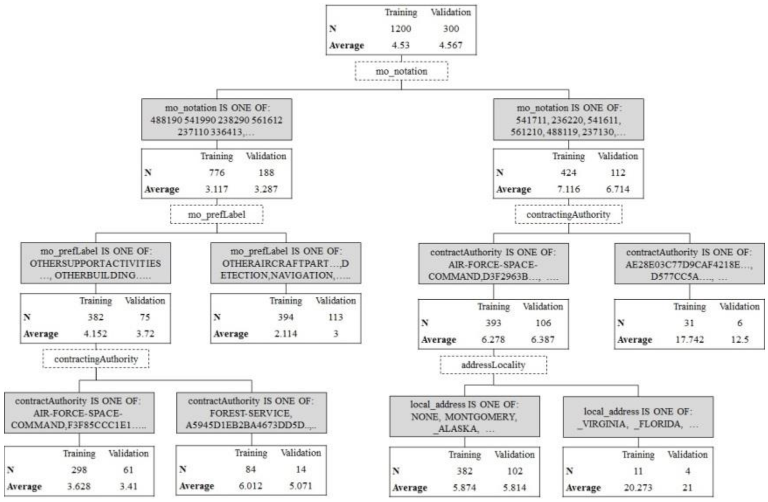


Fig. 3. The subset of Tree 10

IF schema:addressLocality IS ONE OF: _VIRGINIA _FLORIDA _MICHIGAN
AND pc:contractAuthority IS ONE OF:
AIR-FORCE-SPACE-COMMAND F7B62CCDD975E0AD6FCAC227586BFAC3 FBF90D0E606D52317DEB75F84C501BEE
D3F2963B9AE75683B14C6F5F713D5EA8 F3F85CCC1E1008A84317A67C7B534A4E E72630DCCF5374FFB525F42F289A5C23
WASHINGTON-HEADQUARTERS-SERVICES FOREST-SERVICE DLA-ACQUISITION-LOCATIONS AAFD395293B454A34C5089AD694564A0
F13EA6C8864968BD3F1AF2CF4A9E0C84 OVERSEAS-MISSIONS AD6644FFB87D5CD1F2E8F1A87BB87F1F AIR-FORCE-RESERVE-COMMAND
E68F9688F7E463238898EACA8E6CE4EF PACIFIC-AIR-FORCES AIR-COMBAT-COMMAND C60F4CEA542CCCC8691C5AD4C78E10C3
EAD48F3B1E9B8FD8000A32AD745FFBB3 UNITED-STATES-AIR-FORCE-INSTALLA E47F143B8BF6F3BD131CE396EB5B7368
DIVISION-OF-ACQUISITION-AND-COOP AD915CCDFDDB5DBD4F64D0FB46E78347 JOHNSON-SPACE-CENTER FEDERAL-LOCATIONS
DEFENSE-SECURITY-COOPERATION-AGE A25958A5C27643B96776D2267C1A2525 E120C55325D0D737A6A90861EA702FC2
F3B40A5D283E8171DE04F65EB585FA6B EA4F6AE4965B924B9D8C387DAD98B81B AIR-MOBILITY-COMMAND
OFFICE-OF-THE-CHIEF-PROCUREMENT- AF89260E39CAAB41E52DE26A5C1BCA7A D938F7315CD6FF582208AE6B357A086B
E3E9D91BC651BB221A1878D547530CC0 DIRECT-REPORTING-UNITS F41BD406A17168B4B3E0D00007D7A66C
A26EFA8C2BC8AA29AE1A1E8226C9E724 ANIMAL-AND-PLANT-HEALTH-INSPECTI EBA1785813571D306A01C4FEE82B0F1F
AND skos:mo_notation IS ONE OF: 541711 236220 541611 561210 488119 237130 236115
333314 236210 493190 722310 339991 484220 561720 541310 115310 541618
561730 611512 518210 541614 541712 561611 611710 481212 562991 334513
621399 339113 621111 238140 424470 512110 325413 236118 423210 811111
926140 53 19 332993 493110 334111 541410 38 561110

THEN
NODE : 13 / N : 11 / AVE : 20.2727 / SD : 9.9644

Fig. 4. An example result decision rule of Tree 10

4 Conclusion

We have introduced the development procedure of decision tree models to solve the prediction problem of Task A1 of the Linked Data Mining Challenge. The learning procedure of decision trees is performed on the SAS E-miner. Input attributes for learning decision tree algorithm are defined based on the structural information of Public Contracts Ontology. Since the type of target attribute is ordinal non-negative integer number, the variance reduction is used for the attribute selection criterion of decision tree.

One of the limitations of our suggested approach is that input attributes are selected manually, which is inefficient and complicate process when the base data is Linked data. Likewise previously researched decision tree algorithms for linked data [1,4,5], input attributes are needed to be searched automatically through traversing the schema of ontology, even the type of target attribute is ordinal.

References

1. Baader, F., Sattler, U.: An Overview of Tableau Algorithms for Description Logics., Studia Logica, vol. 69, pp. 5-40. (2001)
2. Horrocks, I., Patel-Schneider, P.F.:Reducing OWL Entailment to Description Logic Satisfiability. Re-search Paper at 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Sundial Resort, Sanibel Island, Florida, USA (2003)
3. Jena, http://jena.sourceforge.net/
4. Lehmann, J., Hitzler, P.: Concept learning in description logics using refinement operators. Machine Learning, vol. 78(1-2), pp. 203-250. (2010)
5. Fanizzi, N., d'Amato, C., Esposito, F.: Induction of concepts in web ontologies through terminological decision trees. Proceeding ECML PKDD'10 Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part I, pp. 442-457. (2010)
6. Quinlan, J. R.: Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), pp. 81-106. (1986)