Classification Process of Applicant's data in a Job Agency using Rule-Based Classification Method

Khine Pyae Sone Lwin, May Lwin Than
Computer University, (Pyay)
wankhet14@gmail.com, pontpont.ucsp@gmail.com, cupyay2009@gmail.com

Abstract

Nowadays, everybody has to work and so, they search the associated job balancing with their qualifications. For every graduated people have many opportunities in every workplace. But, they may not be free to search one job after other themselves. So, we need job agency which helps employees find work. In this paper, we propose a job agency system. The goal of this paper is to know the job which concerned with employees' qualifications. This system includes two parts including Training Process which apply pre-defined post and Classification process to classify the incoming employees' qualifications using Rule-Based Classification in Data Mining. This method is to extract rules and then predict unknown class labels (posts). Decision tree can easily convert to classification IF-THEN rules by using decision tree induction. There, ID3 (Iterative Dichotomiser 3) Algorithm will be used to generate decision tree.

1. Introduction

If a job candidate is looking for perfect new job, or an employer is trying to recruit personnel who are reliable and trustworthy then they need to select a job agency with great care, selecting people who carefully consider employer's requirements and put employee forward for the right job based on personal skill and attributes. The last thing, the employee want to go along and meet an employer who wants someone with a particular type of business or technical skill , or that employee's skill base does not match their expectations. So, to ensure everything goes smoothly with job selection process, they need job agency.

There are various types of job. So, classification method is required to select appropriate job. Crucial information should include to have the best chance of success such as qualification, certificate and work experience should be starting point. This system is helping recruit candidates in a huge variety of fields, including accountancy, Information technology, marketing, education, HR and admin, etc.

In the system, Rule-Based classification approach can be used to classify jobs into appropriate posts. In order to improve this system, qualifications are categorized in the training dataset and preprocess the associated posts and the system use ID3 (Iterative Dichotomiser3) algorithm to generate a decision tree that categorize the qualifications as their appropriate post.

2. Related Work

2.1. Classification

Classification, the separation of data records into distinct classes, is apparently the most common data mining task, and decision tree classifiers are perhaps the most popular classification technique.

In this system, Data Classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes, training dataset. The model is represented as classification rules, decision trees, or mathematical formulae. So, the training dataset is used in this system to classify.

The second step is Model Usage. There, accuracy of the model is estimated. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test sets that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur [4, 8]. If the accuracy of the model is considered acceptable, the model can be used to classify future data samples for which the class label is not known.

3. Background Theory

3.1. ID3 Algorithm

Decision trees are one of the most popular methods used for inductive inference. The basic tree induction is a greedy algorithm that constructs decision tree from the top down, with no backtracking [6].

To summarize the ID3 algorithm:

1. For each attribute, compute its entropy with respect to the conclusion.
2. Divide the data into separate sets so that within a set.
3. Build a tree with branches.
4. For each subtree, repeat this process from step- 1.
5. At each iteration, one attribute gets removed from consideration. The process stops when there are no attributes left to consider, or when all the data being considered in a subtree have the same value for the conclusion[7].

3.2. Attribute Selection Measure or Information Gain Measure

The information gain measure is use to select the test attribute at each node in the tree. Such in measure is referred to as *an attribute selection measure or a measure of the goodness of split*. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the lest randomness or "impurity" in these needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

The information gain is based on the decrease in entropy after a dataset is split on an attribute. First the entropy of the total dataset is calculated. The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy. The attribute that yields the largest Gain is chosen for the decision node.

Let S be a set consisting of *s* data samples. Suppose the class label attribute has *m* distinct values defining m distinct classes, C_i (for i=1,…,m). Let s_i be the number of samples of S in class C_i. The expected information needed to classify a given sample is given by

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i) \ldots (1)$$

Where p_i is the probability that an arbitrary sample belongs to class C_i and is estimated by s_i / s . Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have v distinct values, { a_1, a_2, ……, a_v}. Attribute A can be used to partition S into v subsets, {S_1, S_2, ……, S_v}, where S_j contains those samples in S that have value a_j of A. If A were selected as the test attribute (i.e., the have best attribute for splitting), then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots\ldots + s_{mj}}{s} I(S_{1j}, \ldots\ldots\ldots\ldots, S_{mj}) \ldots (2)$$

The term $\frac{s_{1j} + \ldots\ldots + s_{mj}}{s}$ acts as the weight of j^{th} subsets and is the number of samples in the subset (i.e., having value a_j of A) divided by the total number of samples in S. The smaller the entropy value, the greater the purity of the subset partitions. Note that for a given subset S_j.

$$I(S_{1j}, S_{2j}, \ldots\ldots\ldots S_{mj}) = -\sum_{i=1}^{m} p_i \log_2 (p_i) \ldots (3)$$

Where $p_{ij} = \frac{s_{ij}}{|S_j|}$ and is the probability that a sample in S_j belongs to class C_i .The encoding information that would be gained by branching on A is

$$\text{Gain}(A) = I(S_1, S_2, \ldots\ldots\ldots, S_m) - E(A) \ldots\ldots (4)$$

The other words, Gain(A) is the expected reduction in entropy caused by knowing the value of attribute A. The attribute with the highest gain ratio is chosen as the test attribute for the given set S.

3.3. Classification Accuracy

In classification, each training sample can belong to only one class and it is commonly assumed that objects are uniquely classifiable. Classification algorithms can be compared according to their accuracy. Accuracy is measured using a test set of objects for which the class labels are known.
Accuracy is estimated as the number of correct class predictions. And, it is divided by the total number of test samples.

4. System Design

In this figure, data classification is a two-step process. In the first step, a classifier is built by describing a predetermined set of data classes. A classification algorithm builds the classifier by analyzing or learning from a training data of database tuples and their associated class labels. In this system, ID3 algorithm is used to classify the applicant's data.

In the second step, Rule-based classifier is built by extracting IF-THEN rules from a decision tree. To extract rules from a decision tree, one rule is created for each path from the root to a leaf node. Finally, a new job application from is used to classify the class label with predetermined classes in job agency.

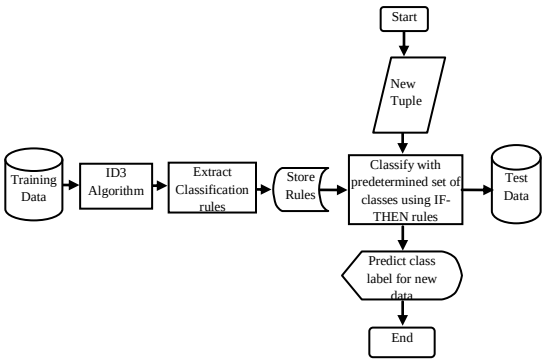


Figure (1). System Flow Diagram

Table (1). Sample Training Dataset

Qualification	***Certificate***	***Experience***	***Age***	***Language***	***Post***
Any Graduate	LCCI Level3	Under 3 years	over 30	English	Junior Accountant
Any Graduate	MYOB	Under 3 years	over 30	English	Junior Accountant
Any Graduate	ACCA	Over 3 years	under 30	Korean, English	Senior Accountant
B.B.A	LCCI Level3	Under 3 years	over 30	English	Senior Accountant
Any Graduate	Computer Basic, A.B.E	Under 3 years	over 30	English	Junior Admin
Any Graduate	Computer Basic, A.B.E, M.HR	Under 3 years	over 30	English	Junior Admin
Any Graduate	Computer Basic, A.B.E, M.HR	Over 3 years	under 30	Korean, English	Senior Admin
Any Graduate	Computer Basic, A.B.E	Over 3 years	over 30	Japanese, English	Senior Admin
Any Graduate	Computer Basic, A.B.E, M.HR	Under 3 years	over 30	English	Junior Web Designer

Any Graduate	Computer Basic, A.B.E, M.HR	Under 3 years	over 30	English, Chinese	Junior Web Designer
Any Graduate	Computer Basic, A.B.E, M.HR	Over 3 years	over 30	English	Senior Web Designer
Any Graduate	Computer Basic, A.B.E, M.HR	Over 3 years	over 30	English, Chinese	Senior Web Designer
Any Graduate	Computer Basic, A.B.E, M.HR	Under 3 years	under 30	Japanese, English	Junior Network Engineer
Any Graduate	Computer Basic, A.B.E, M.HR	Under 3 years	under 30	Korean, English	Junior Network Engineer
Any Graduate	Computer Basic, A.B.E, M.HR	Over 3 years	over 30	English	Senior Network Engineer
Any Graduate	Computer Basic, A.B.E, M.HR	Over 3 years	over 30	English, Chinese	Senior Network Engineer
Any Graduate	Photoshop , Corel Draw	Over 3 years	over 30	Korean, English	Senior Designer
Any Graduate	Photoshop , Corel Draw	Over 3 years	under 30	English	Senior Designer
Any Graduate	Photoshop , Corel Draw	Under 3 years	over 30	English	Junior Designer
Any Graduate	Photoshop , Corel Draw	Under 3 years	over 30	English, Chinese	Junior Designer

Table (2). Attribute and Values of Proposed System

Attribute	***Description***
Qualification	Any Graduate, B.A(Eco.), B.B.A, B.Com., M.B.A
Certificate	ACCA, CCNA, (Computer Basic, A.B.E), (Computer Basic, A.B.E, D.M.A), (Computer Basic, A.B.E, M.HR), (Computer Basic, D.M.A), (Computer Basic, D.M.A, A.B.E, M.HR), (Computer Basic, D.M.A, M.HR), (Computer Basic, M.HR), (HTML, Web), LCCI Level3, MYOB, (Photoshop, CorelDraw, AutoCad), (Photoshop, CorelDraw, AutoCad, 3D MAX)
Language	English, (English, Chinese), (Japanese, English), (Korean, English)
Experience	Over 3 years, Under 3 years
Age	over 3, under 3

5. Attribute Selection by Information Gain Computation

Information Gain is used to select best attribute. Entropy for experience is computed:

D=Dataset

Info(D)= 0.7492

E(experience)= 0.4722

And Then,

Gain(Experience)=Info(D)-E(experience)

= 0.7492- 0.4722

= 0.2770

Similarly,

Gain(Certificate)=1.7168

Gain(Qualification)=0.8500

Gain(Age)=0

Gain(Language)=0

Certificate has the highest information gain among attributes, it is selected as these splitting attribute. So, Certificate is defined as root node and branches are grown for each of the attribute's values.

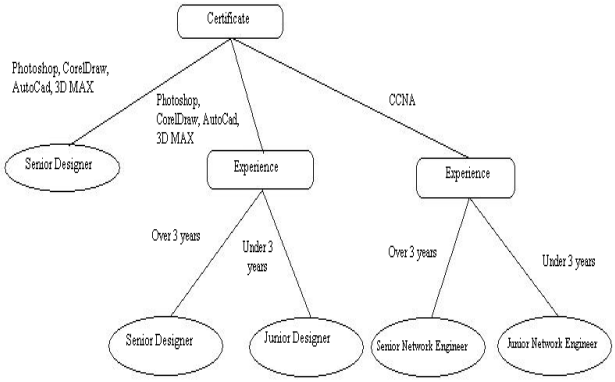
5.1. Sample Decision Tree

5.2. Extracted Rules from Decision Tree

Training

MainNode	ChildNode	ParentNode	MainParent	Result
Certificate	ACCA	ROOT	Start	
ACCA	Qualification	Certificate	ROOT	
Qualification	Any Graduate	ACCA	Certificate	
Any Graduate	Experience	Qualificaton	ACCA	
Experience	Over 3 years	Any Graduate	Qualificaton	
Over 3 years		Experience	Any Graduate	Senior Accountant
Experience	Under 3 years	Any Graduate	Qualificaton	
Under 3 years		Experience	Any Graduate	Junior Accountant
Qualificaton	B.A(Eco)	ACCA	Certificate	
B.A(Eco.)	Experience	Qualificaton	ACCA	
Experience	Over 3 years	B.A(Eco.)	Qualificaton	
Over 3 years		Experience	B.A(Eco.)	Senior Accountant
Experience	Under 3 years	B.A(Eco.)	Qualificaton	
Under 3 years		Experience	B.A(Eco.)	Junior Accountant
Qualificaton	B.B.A	ACCA	Certificate	
B.B.A		Qualificaton	ACCA	Senior Accountant
Qualificaton	B.Com.	ACCA	Certificate	
B.Com.		Qualificaton	ACCA	Senior Accountant
Qualificaton	M.B.A	ACCA	Certificate	
M.B.A		Qualificaton	ACCA	Senior Accountant
Certificate	CCNA	ROOT	Start	
CCNA	Experience	Certificate	ROOT	
Experience	Over 3 years	CCNA	Certificate	
Over 3 years		Experience	CCNA	Senior Network Engineer
Experience	Under 3 years	CCNA	Certificate	
Under 3 years		Experience	CCNA	Junior Network Engineer
Certificate	Computer Basic, A.B.E	ROOT	Start	
Computer Basic, A.B.E	Experience	Certificate	ROOT	
Experience	Over 3 years	Computer Basic, A.E.E	Certificate	
Over 3 years		Experience	Computer Basic, A.B.E	Senior Admin
Experience	Under 3 years	Computer Basic, A.E.E	Certificate	
Under 3 years		Experience	Computer Basic, A.B.E	Junior Admin
Certificate	Computer Basic, A.B.E, D.M.A	ROOT	Start	
Computer Basic, A.B.E, D.M.A		Certificate	ROOT	Senior Admin

5.3. Decision Tree for Prospect Level



5.4. Rules Extraction from Sample Dataset

As seen from the decision tree in Figure, there are a total of five paths in the tree, indicating that five Classification Rules can be extracted, which are stated below. These rules will be used for prediction in next step.

R1: If Certificate=(Photoshop, CorelDraw, AutoCad, 3D MAX)
then Post=Senior Designer

R2: If Certificate = (Photoshop, CorelDraw, AutoCad) AND Experience = Over 3 years
then Post= Senior Designer

R3: If Certificate = (Photoshop, CorelDraw, AutoCad) AND Experience = Under 3 years
then Post= Junior Designer

R4: If Certificate = CCNA AND Experience = Over 3 years
then Post= Senior Network Engineer

R5: If Certificate = CCNA AND Experience = Under 3 years
then Post = Junior Network Engineer

5.5. Classification Accuracy

In this proposed system, classification accuracy is computed with hold out method. The hold out method reserves a certain amount of data for testing and uses the remainder for training-so they are disjoint. To generate classification model, training dataset is used. A test sample is used for the resulting classification model in terms of accuracy.

The accuracy is computed:

$$\text{Accuracy} = \frac{numbers\ of\ correct\ count \times 100}{numbers\ of\ total\ records}$$

Example:
Testing Rules(testing record #1)= record #1.class - Succ
Testing Rules (testing record #2) not= record #2.class - Error
Testing Rules (testing record #3) = record #3.class - Succ
Testing Rules (testing record #4) = record #4.class - Succ
Testing Rules (testing record #5) not= record #5.class - Error
Error rate:
2 errors: #2 and #5
Correct Record = Total records- Error Records
= 5-2
=3
Accuracy: 3/5 × 100 = 60%

6. Conclusion

This paper presents the implementation for user to view the knowledge of data mining. And, nowadays computer job announcement posting system has been applied in rapid action in worldwide countries. Therefore the development of computer job announcement posting system in Data Mining is useful for everyone who liked to check their qualification. The users can choose their qualification from the system that shows the available job type. It supports to decision makers from middle management upward with information at the correct level of detail to support decision-making and without time consuming. The database of the proposed system is used MS.Excel and C#.Net is used. The system is useful for the users who want to know their qualification fit with it and without searching one job after other themselves.

7. Further Extensions

In this module, some extensions are proposed to increase the capabilities and efficiency of Data Mining System. The system has been tested on job announcement posting dataset. We can use any other dataset for jobs cases using the decision tree algorithm. We are going to evaluate the accuracy by using different methods.

8. References

[1] Aijun, A, "Classification Methods" York University, Canade
[2] F. Berzal and Nicolas Marin, "Data Mining Concepts and Techniques"
[3] Galant Violetta, Owoc Mieczyslaw L, Gladysz Tomasz, "Applying Decision Trees in Classification Tasks"
[4] Han, J and Kanber, M, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001
[5] Quinlan J.R. "Induction of Decision Trees. Machine Learning" 1986/1
[6] Allan Neymark. CS157B – Spring 2007. Agenda. "Decision Trees; What is *ID3*?"
[7] Rule induction-Edinburgh **Napier** University, Peter Ross, x4437 This version: 2000-10-30
[8] Win Mar Oo, "Biomedical Data Ananysis for Diabetes using Hybrid Learning Method with Genetic Algorithms and Decision Tree", University of Computer Studies, Yangon
[9] Min Lwin, "Classification of Military officer's Rank Promotion using Decision Tree Algorithm", University of Computer Studies, Yangon
[10] Nan New Ni Tun, "Development of Rule Induction System Using Decision Tree Algorithm for Computer Job Annonment Posting", University of Computer Studies, Yangon
[11] "A Rule-Based Classification Algorithm for Uncertain Data", Sunil Prabhakar, Department of Computer Science, Purdue niversity, sunil@cs.purdue.edu,