A Probabilistic Machine Learning Approach for Eligible Candidate Selection

Marium-E-Jannat
Department of Computer Science and Engineering
Shahjalal University of Science and Technology Sylhet, Bangladesh

Sayma Sultana Chowdhury
Department of Computer Science and Engineering
Shahjalal University of Science and TechnologySylhet, Bangladesh

Munira Akther
Department of Computer Science and Engineering
Shahjalal University of Science and Technology Sylhet, Bangladesh

ABSTRACT
Now-a-days Machine learning approach is used to solve many problems where intelligence is involved. Lots of time consuming task are done by computers with the power of statistics. In this paper, a machine learning based candidate selection procedure is proposed and implemented for a particular field. A huge amount of activity is involved in the job recruitment procedure. To reduce the manual task a probabilistic machine learning approach is described in this paper. A popular machine learning approach named Naive Bayes Classifier is used to implement the method. Baseline criteria selection depends on the recruiters demand. The proposed system learns from training dataset and produces a short listed eligible list based on learning. The more perfectly one feed the system result will be more accurate.

General Terms
Candidate Selection, Job Recruitment.

Keywords
Machine Learning, Probability, Statistics, Naive Bayes Classifier.

1. INTRODUCTION
We are living in the most technology oriented time. Day by day technology reduces human task. Human are interested in automation and wants to reduce the burden of manual task. As it is the most competitive world, employers need to handle lots of applicants for job interview. So it is a very natural need to automate the selection procedure. Focusing on this fact, this paper presents the plan to establish the system which will help to make a short list of applicant. Usually software firms recruit people through viva or written exam or skill test. They come to campus or call to their office with Curriculum Vitae, Which contains information about CGPA, projects, interest, research work. A matured firm can take decision from their experience. A startup firm can take decision from any well established company if they want. Based on this concept, the system learns from previous experience of a well established company. They have freedom to change criteria or select criteria according to their company demand.

2. PROBLEM DESCRIPTION
Real life scenario of the competitive world is that there are different types of students and academic performance and result do not always define them or their skills. So only CGPA can never define them as they will be good employee or not. Different company need specialized person in different sector. To solve this problem the Probabilistic Machine Learning Approach for Eligible Candidate Selection system is designed for software firm. The training set enrich with the data of previous experience of well established software firm. Using the training set a machine learning approach is implemented. Machine learning works on previous data and predicts on the base of whatever it learns. Since the system learns from training data set so it is a supervised learning system. Parameter need to be selected by recruiter. Base on software firm total eleven parameters are selected for this system.

3. BACKGROUND STUDY
The primary motto was to develop such a system which will reduce the load of manual paper work of different domain. So, there was a try to find out a solution to give intelligence to machine so that it can help to make decision by predicting the best people for a specific domain.

When going through different types of machine learning projects/research work for predicting something in machine learning approach; a project work of Stanford University, titled "A Machine Learning Approach for Future Career Planning" [1] make us interested for this particular research

The above mentioned paper titled "A Machine Learning Approach for Future Career Planning" by Yu Lou, Ran Ren and Yiyang Zhao from Stanford University states a way of modeling of peoples career choices through a machine learning approach. In their approach, they collected profiles of a large number of people and extracted features from those profiles' descriptive information. They used Markov Chain as machine learning tool to estimate the transition probability matrix. Based on this probability calculation, they recommend a person the best career development pathway according to his/her current goal and status. [2]

In another paper titled "Feature selection for unbalanced class distribution and Naive Bayes" [3]; Dunja Mladenic and Marko Grobelnik describes a way to select subset of feature and developed the Naive Bayesian classifier based on that feature set to apply on text data. They concentrated mainly on domains with many features having a highly unbalanced class distribution and asymmetric misclassification costs. By the term, asymmetric misclassification costs they meant the class value of target class for which the system will give predictions. Their prescribed method prefers false positive over false negative.

4. METHODOLOGY

4.1 Proposed Method

In Machine learning, there are supervised and unsupervised learning approaches are used. A supervised learning approach learn from training dataset, analyzed them and take decision based on training data. Naturally previous experiences are used in employee recruitment procedure. So supervised machine learning approach is selected for this system design. One of the most powerful supervised Machine learning approaches is Naive Bayes classifier. Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.[wiki] In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on conditions that might be related to the event. Bayes' theorem is stated mathematically as the following equation: [4]

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

Where *A* and *B* are events and P(B) ≠ 0.

- *P*(*A*) and *P*(*B*) are the probabilities of *A* and *B* without regard to each other.
- *P*(*A* | *B*), a conditional probability, is the probability of observing event *A* given that *B* is true.
- *P*(*B* | *A*) is the probability of observing event *B* given that *A* is true.

$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n)$ The most probable class using Bayes Rule is v_{MAP}

$$= \arg\max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n \mid v_j) P(v_j)}{P(a_1, a_2, ..., a_n)}$$

$$= \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n \mid v_j) P(v_j)$$

The system is proposed to be implemented according to Naive Bayes Classifier.

4.2 Dataset

For training dataset data can be gathered from a reputed software firm. Parameters should be defined. Rank level depends on recruiter's choice. For this Naive Bayes based system parameters are ranked into following levels:

People with very poor knowledge in respective attribute – Level 1

People with poor knowledge in respective attribute – Level 2

People with medium knowledge in respective attribute – Level 3

People with high knowledge in respective attribute – Level 4

Following parameters are used: GPA in C course, GPA in Java, Algorithm, Data Structure, Project work, Thesis work, Total CGPA, Extra Curricular Activities, and Database Management. System Design and implementation procedure accomplished based on this dataset.

4.3 Implementation

For this particular task, a set of hypothesis from real life experience was collected where some people were seen, who have enough CGPA, enough knowledge in programming language but not well behavior or not interested to work it would be a great thread to maintain firm's environment. In real life experience also see total CGPA is not enough but programming skill is very good also enough interest and he/she do well in job and firm also benefited from them. The dataset have a set of hypothesis which is collected from real life experiment. After calculating frequency and probability, Naive Bayes theorem was applied to compute most probable class and able to take decision about someone. This paper shows a test case having high CGPA (4) but not eligible for job on the other hand, a test case having low CGPA (3) but eligible for job considering other parameter.

Table 1: A chunk of training data set of the classification task

Sample	C language	Java	Algorithm	Data Structure	Database	Attitude	Project	Thesis	Software Engineering	Interest	Curricular Activities	Total CGPA	Class
Person-1	1	3	2	2	3	2	2	1	3	4	3	1	**no**
Person-2	4	2	4	4	2	3	4	3	2	4	1	4	**yes**
Person-3	4	4	1	3	3	4	3	4	4	1	3	3	**yes**
Person-4	3	1	3	1	4	4	2	3	3	3	2	4	**yes**
Person-5	3	4	3	3	3	1	4	4	1	4	1	4	**no**
Person-6	4	2	2	4	1	1	1	2	3	3	3	3	**no**
Person-7	1	4	4	4	4	4	1	2	4	2	2	1	**yes**
Person-8	3	3	2	2	4	1	4	4	1	4	4	3	**yes**
Person-9	2	4	4	3	1	2	4	1	3	4	1	2	**yes**
Person-10	3	3	4	1	4	4	3	3	4	1	4	4	**no**
Person-11	2	1	1	2	2	3	2	2	2	2	2	2	**no**

Table 2: Probabilities of above mentioned chunk of training dataset

	Class	Level 1	Level 2	Level 3	Level 4
C Language	yes	1/6	1/6	2/6	2/6
	no	1/5	1/5	2/5	1/5
Java	yes	1/6	1/6	1/6	3/6
	no	1/5	1/5	2/5	1/5
Algorithm	yes	1/6	1/6	1/6	3/6
	no	1/5	2/5	1/5	1/5

Data structure	yes	1/6	1/6	2/6	2/6
	no	1/5	2/5	1/5	1/5
Database	yes	1/6	1/6	1/6	3/6
	no	1/5	1/5	2/5	1/5
Attitude	yes	1/6	1/6	1/6	3/6
	no	2/5	1/5	1/5	1/5
Project	yes	1/6	1/6	1/6	3/6
	no	1/5	2/5	1/5	1/5
Thesis	yes	1/6	1/6	2/6	2/6
	no	1/5	2/5	1/5	1/5
Software Engineering	yes	1/6	1/6	2/6	2/6
	no	1/5	1/5	2/5	1/5
Interest	yes	1/6	1/6	1/6	3/6
	no	1/5	1/5	1/5	2/5
Curricular activities	yes	2/6	2/6	1/6	1/6
	no	1/5	1/5	2/5	1/5
Total CGPA	yes	1/6	1/6	2/6	2/6
	no	1/5	1/5	1/5	2/5
Class	yes	6/11			
	no	5/11			

Table 3: Sample test case

C language	Java	Algorithm	Data Structure	Database	Attitude	Project	Thesis	Software Engineering	Interest	Curricular Activities	Total CGPA
3	3	2	2	3	4	3	3	3	4	4	4

$V_{NB\ (yes)} = 6/11*2/6*1/6*1/6*1/6*1/6*3/6*1/6*2/6*2/6*3/6*1/6*2/6 = 3.608328368e\text{-}08$

$V_{NB\ (no)} = 5/11*2/5*2/5*2/5*2/5*2/5*1/5*1/5*1/5*2/5*2/5*1/5*2/5 = 4.766254545e\text{-}07$

$V_{NB\ (yes)} < V_{NB\ (no)}$

So, V_{NB} = NO (not eligible)

4.4 Result Analysis

A test implementation of the system was done for a software firm where input for employee was given according to the system. A test case is given below.

C Programming Language: 4
Java Programming Language: 3
Algorithm: 3
Data Structure: 2
Database: 3
Software Engineering: 3
Project: 3
Thesis: 3
Total CGPA: 3
Attitude: 4
Interest: 3
Extra Curricular Activities: 1
Result= YES (eligible)

Using this method companies can easily make a short list of candidates automatically. According to the choice of any company the system produce result accurately.

5. PSEUDO CODE

System is implemented and executed. The Pseudo code is given bellow.

$$Naive\_Bayes\_Learn(examples)$$
$$For\ each\ target\ value\ v_j$$
$$\hat{P}(v_j) \leftarrow estimate\ P(v_j)$$
$$for\ each\ attribute\ value\ a_i\ of\ each\ attribute\ a$$
$$\hat{P}(a_i \mid v_j) \leftarrow estimate\ P(a_i \mid v_j)$$
$$Classify\_New\_Instance(x)$$
$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i \mid v_j)$$

6. CONCLUSION

A Probabilistic Machine Learning Approach for Eligible Candidate Selection is a unique method for shortlisted eligible candidates. It performs accurately based on training dataset. As the training dataset contain input from previous experiment so the system's performance closely to human being. It is the matter of hope that this system is successful to reduce the huge amount of manual data manipulation. This system is able to solve the candidate selection problem in an intelligent manner using statistics. So this powerful machine learning approach should be implemented in all sectors. Based on current scenario anyone can try another machine learning approaches to be implemented in future for this particular problem. By measuring performance among different machine learning or statistical approaches, the most effective approach can be decided which will do the task most perfectly.

7. REFERENCES

[1] Yu Lou, Ran Ren and Yiyang Zhao; Stanford University; "A Machine Learning Approach for Future Career Planning"

[2] http://cs229.stanford.edu/proj2010/LouRenZhang-AMachineLearningApproachForFutureCareerPlanning.pdf ; Access Date: 30th April, 2016.

[3] Dunja Mladenic and Marko Grobelnik; J. Stefan University; "Feature selection for unbalanced class distribution and Naive Bayes"; Proceedings of the 16th International Conference on Machine Learning (ICML).

[4] https://en.wikipedia.org/wiki/Bayes%27_theorem; Access Date: 8th June, 2016.

[5] [Gedikli et al. 2011] RF-Rec: Fast and accurate computation of recommendations based on rating frequencies, Proceedings of the 13th IEEE Conference on Commerce and Enterprise Computing - CEC 2011, Luxembourg, 2011.

[6] William Feller, "An Introduction to Probability Theory and Its Applications", Volume 1, 3rd Edition, ISBN: 978-0-471-25708-0, January 1968.

[7] John O'Donovan & Barry Smyth; "Trust in Recommender Systems"; Proceedings of the 10th international conference on Intelligent user interfaces; Pages 167-174, ACM, New York, NY, USA; 2005; ISBN:1-58113-894-6.

[8] J. Ben Schafer, Joseph Konstan, John Riedl ; "Recommender Systems in E-Commerce"; Proceedings of the 1st ACM conference on Electronic commerce Pages 158-166, ACM, New York, NY, USA, 1999, ISBN:1-58113-176-3.

[9] Mike Oaksford, Birkbeck College, London; Nick Chater, University College London; "Bayesian Rationality: The Probabilistic Approach to Human Reasoning"

[10] Haun Liu, Rudy Setiono; National University of Singapore; "A probabilistic approach to feature selection-a filter solution", ICML, 1996 – Citeseer

[11] S.B. Kotsiantis, University of Peloponnese, Greece; "Supervised Machine Learning: A Review of Classification Techniques"; Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press, 2007.

[12] Joachim Herbst, "A Machhine Learning Approach to Workflow Management"; 11th European Conference on Machine Learning Barcelona, Catalonia, Spain, May 31 – June 2, 2000; Online ISBN: 978-3-540-45164-8

[13] Nava Tintarev, Judith Masthoff; University of Aberdeen, Scotland, U.K; "A Survey of Explanations in Recommender Systems"; Data Engineering Workshop, 2007 IEEE 23rd International Conference on; E-ISBN: 978-1-4244-0832-0; Pages: 1-810

[14] G. Adomavicius ; Carlson Sch. of Manage., Minnesota Univ., Minneapolis, MN, USA ; A. Tuzhilin;Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions Published in: IEEE Transactions on Knowledge and Data Engineering (Volume:17 , Issue: 6 , Pages: 734-749); ISSN: 1041-4347

[15] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, Tao Zhou; "Recommender Systems"; Physics Reports, Volume: 519 Issue: 1, October,2012 Pages: 1-49, Elsevier

[16] Paul Resnick, AT&T, Labs-Research, Murray Hill, N.J.; Hal R. Varian, Dean of the School of Information Management and Systems, University of California, Berkeley; "Recommender Systems"; Communications of the ACM, Volume 40 Issue 3, March 1997, Pages 56-58, ACM, New York, NY, USA.