Making Better Job Hiring Decisions using "Human in the Loop" Techniques

Christopher G. Harris

University of Northern Colorado, Greeley, CO 80639 USA
christopher.harris@unco.edu

Abstract. Using machine learning techniques to filter and sort job candidates has been done for more than two decades; however, there are always humans involved in the final hiring decision. One primary reason is that rarely are two hiring decisions made with the same information and in the same context. Many experts believe that any information that can be passed from one human decision-maker to another can also be passed to a machine. Through empirical experiments, we look at ways in which this human feedback can be used to better train machine learning algorithms with special attention to the inherent risks, such as overfitting data and avoiding bias.

Keywords: Human resources, machine learning, feedback mechanisms, job hiring, artificial intelligence

1 Introduction

Companies have long understood that hiring the best employees produces a competitive advantage which is hard for its competitors to duplicate. One of the more significant challenges in finding the most appropriate applicant for a job listing is its inexact nature - one that is influenced by "feel" as much as by skills and talent.

For over two decades information retrieval systems have been used by human resources (HR) departments and external headhunting services to filter and sort candidates based on a set of weighted features gathered from the cover letters and résumés or curriculum vitae (CVs), interviews with the candidate, letters of recommendation, as well as other supporting materials such as transcripts or certifications held. These systems are becoming ubiquitous; with 74% of large U.S. organizations using some form of electronic selection tool to help with the hiring process [1]. These systems have been able to save considerable time and money in the recruiting process [2].

These retrieval systems combine the tasks of natural language processing (NLP), data and text mining, as well as rule-based logic. More recently, systems have employed artificial intelligence (AI) to identify which candidates are likely to accept the job offer, which are not likely to look for a job with another firm within the first year or two, or which are most likely to move up the ranks into management.

Most machine techniques score candidates based on keyword and phrase matches. Many also apply machine learning algorithms that employ associative rules, classification rules, clustering patterns, and/or prediction rules and patterns. Of these four types of machine learning algorithms, those that use classification rules and prediction rules and patterns to categorize candidates into different groups are used most frequently. For instance, candidates could be grouped as highly suitable, potentially suitable, and not suitable. More advanced systems use Knowledge Discovery in Databases (KDD) to provide more accurate decision support. These decision support systems can look at the performance of other employees and make longitudinal predictions on candidates showing similar traits. Figure 1 illustrates how data can be combined with rules to provide decision rules and data for decision support tools.

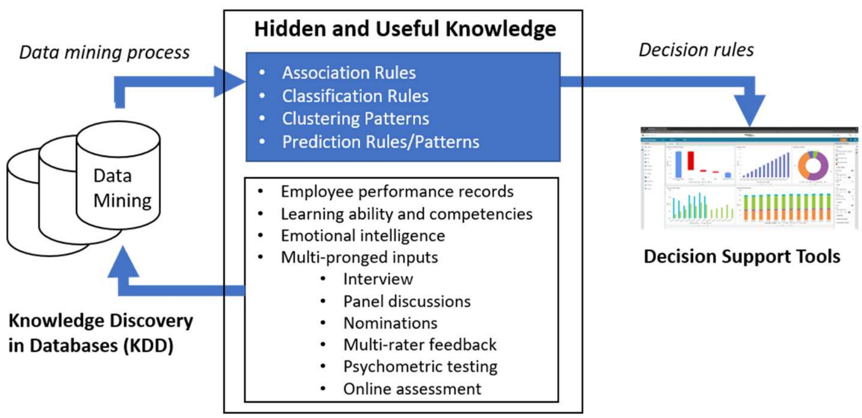

Fig. 1. Illustration of how hidden and useful knowledge can be combined with machine learning techniques to provide inputs to decision support tools. Other types of knowledge can be passed to KDD systems for data mining, allowing better decisions to be made. Adapted from [3].

In data mining tasks, classification and prediction is among the popular task for KDD and making future predictions. The classification process is known as supervised learning, where the classification target is already known. The decision tree technique has its advantages such as it can produce a model which may represent interpretable rules or logic statement; it is more suitable for analyzing categorical outcomes; it is non-parametric which is suited to capture a functional form relating independent and dependent variables; easy to interpret, computationally inexpensive, capable in dealing with noisy data, its prediction model is intuitively explainable to the user, it has automatic interaction detection to find the significant high-order interactions quickly, and it can produce more informative outputs [4][5]. The C4.5 classification algorithm is easy to understand as the derived rules have a very straightforward interpretation. For these reasons, we use the C4.5 classification algorithm in our study.

2 Job Hiring Challenges

One challenge in the job hiring process is that nearly all job searches are unique. Even if two job searches for the same job title in the same department of the same company require the same qualifications and use the same hiring committee, the hiring for each will be done in different contexts. They will draw in completely different candidates, each with their own unique set of strengths and weaknesses. Often a company or department is likely to have different objectives to meet from one quarter to the next. The work teams looking at the "fit" of a new hire are comprised of a different mix of people over time due to other recent hires, job transfers, and employee departures. For this reason and many others, job searches remain difficult for a human to do effectively; for a machine, the task is even more daunting.

Even with the assistance from machines to filter and sort candidates, selecting the most appropriate job applicants for a job position is not only subjective but also requires years of experience to accomplish well. HR and executive recruitment and search firms typically undertake several approaches to select the most appropriate résumés or CVs, such as completing a checklist or rubric for each submitted résumé or CV and performing a keyword search on a collection of application materials. Utilizing human HR experts is generally viewed as the most effective method of search, yet it is not without significant disadvantages – it is resource intensive and does not scale well (i.e., a human expert can only review a limited set of applicants per day). Companies such as Johnson & Johnson, a consumer goods company, receive 1.2 million applications for 25,000 positions annually [6]. At the same time, the more experienced HR and executive search staff often must focus their attention on maintaining corporate accounts or attending to other needs, relegating applicant screening to junior-level employees or outsourcing it to outside firms with far less experience and without a true understanding of the hiring manager's needs.

While these algorithm-based techniques have allowed job searches to utilize better, more focused matching methods to and improve accuracy and consistency, they have not yet been proven reliable enough to entirely remove humans from the search task. For example, they can be reverse engineered by candidates, and semantic analysis using natural language processing (NLP) methods have not advanced to the level that humans innately possess. Also, despite many advances in AI and Information retrieval (IR) systems, hiring decision makers must carefully consider the costs of false positives and false negatives in the hiring process. Overall, during candidate screening, companies try to reduce the number of false negative candidates (the potential superstar employees that the filtering software may reject) at the expense of false positives (the low performers the filtering software does not reject), since in nearly every case we have examined, human judgment will be used to further screen candidates prior to hiring. Therefore, both machine and human-in-the-loop approaches are necessary to reduce the pool of applicants to those who are indeed a suitable match.

In general, there are two categories of skills that job searches hope to ascertain about each candidate. The first is the skills required by the job, whether it be language-specific (e.g., Java, Python, French) or task-specific (e.g., leading a sales team, launching a new brand of clothing, developing full-stack software). However, these approaches do not adequately address the nuances of a great potential employee. The second are soft skills that determine fit, motivation, and attitude. For many management-level jobs, these softer skills are viewed as equally important and is where technology-based search solutions are often challenged and where humans can provide the best guidance [7]. However, there are many who believe an evaluation of these soft skills can be given to a machine (e.g., [8][9]), much in the way a senior HR employee can guide a junior employee in soft skill evaluation of a pool of candidates.

Although there are many AI-based models described in the literature (e.g., [10], [11]), there are few empirical studies involving actual HR data. Strohmeier and Piazza [12] indicate this is likely due to the quality of available data, which limits researcher's ability to conduct empirical studies. One notable exception can be found in [5], Chien and Chen perform a case study to illustrate how a decision tree can aid in selecting personnel for a job in the high technology sector.

In this paper, we conduct an empirical evaluation of how humans in the loop can be used to better train AI systems. Since the algorithm used to match a candidate to a job is not consistent from one job search to the next, the weights and features evaluated by the machine learning algorithm change too, limiting the transfer of information. We seek to find answers to the following research questions:

1. **How consistent are HR experts in determining the best features of a candidate's materials to evaluate?** If there is little to no consistency between experts, it is challenging to develop an algorithm to replicate the human expert. Moreover, it is easy to introduce bias into the hiring process.
2. **How can human experts provide input to better train a machine, particularly on the softer skills?** The job requirements for softer skills are often vaguely written. Also, companies want to make it easy for candidates to apply; therefore, they accept a cover letter and CV/resume only and don't use standardized questionnaire to obtain this information. It is up to the human or machine to determine if the candidate has the required skills from the materials they have provided.

3 Experiment Design

To examine the first research question, we began with the set of 5 job descriptions in English for management-level job positions (see Table 1). We chose management level job descriptions since they are more likely to have a mix of both language specific and soft skills. These job descriptions were taken from actual job searches conducted in 2016-17. We followed many of the anonymity procedures for both candidates and companies as mentioned in [13] and [14]. For instance, to avoid potential bias, information about the hiring company was removed or made generic to avoid identifying prospective employers. We removed those job applicants from each pool that did not meet the minimum job requirements for experience, education, or job location listed in the job listing. After removing applicants that did not meet the minimum job requirements, a sizeable pool of applicants for each job description remained (M=58.6, SD=13.7).

Table 1. Job titles for 5 actual job positions, total number of applicants and number of applicants remaining after screening for minimum qualifications, as used in our study

Job Title	Job Location	Total # of Applicants	# of Applicants Remaining After Screening
1. Assistant Manager/ Technical Supervisor	Hong Kong	92	64
2. Manager, Project & Network	Singapore	73	61
3. Senior Manager, IT Mgt & Service Integration	New Jersey, USA	48	39
4. Unit Manager / Business Development Manager	Guangzhou, China	106	76
5. Operations Manager	California, USA	79	53

From each of the 5 pools of applicants, 20 applicants were randomly selected from the pool of actual received submissions. All non-standard acronyms in each job posting's description and in the job application information was resolved (expanded) for clarity. All personally-identifying data were obscured or genericized to make all candidates and companies non-identifiable.

To answer our first research question, we asked a group of 13 HR personnel (average number of years of experience in HR = 9.0) to weigh which of the identified features from the pool of 100 résumés and cover letters were most relevant to make hiring decisions. These features and their relative importance are important inputs to the machine learning algorithms. A list of the 10 features that experts ranked highest is provided in Table 2.

From Table 2, we can see the ranking of features differs greatly between experts. To represent this numerically, we use rank-biased overlap (RBO) as our metric [15]. RBO has several important advantages over the more commonly-used Kendall Tau, namely it does not suffer from the disjointedness problem (when a job candidate appears in one ranked list but not another) and RBO weighs those that match towards the top of the ranked list more heavily than those that match toward the bottom – two properties the Kendall Tau metric does not possess. RBO is measured on a scale of 0 (completely disjoint) to 1 (a perfect match). We obtained an RBO score of 0.189 and 0.215 for the top 10 and top 5 respectively, indicating little evidence of ranking consensus.

Table 2. Ranking of features from job candidate materials, the number of appearances in top 10 lists and top 5 lists of importance, as determined by our 13 HR experts.

Rank	Feature	# in top 10	# in top 5	Rank	Feature	# in top 10	# in top 5
1.	Years of relevant work experience	7	4	6.	No notable gaps in employment	4	3
2.	Job responsibilities held	6	3	6.	Salary expectations	4	3
3.	Technical skills match requirements	6	2	8.	University attended	3	2
4.	Education level attained	5	3	9.	Job titles held	3	1
5.	Job promotions earned	4	2	10.	Languages spoken	2	1

In addition, to examine if there was some form of possible agreement on features possible between our human experts, we presented all with the overall ranking from Table 2. We asked if this ranking, converted into a score, could be used for the pool of candidates. This provided considerable discussion between them with a majority (9 of the 13) disagreeing with the utility of using the features in the rank order that was collaboratively determined. Most came up with specific examples why this ranking of features would not work from the pool of 100 resumes they examined.

There are several important implications from this. First, it makes training a machine learning algorithm to screen and select job candidates challenging, since human experts are the oracle these algorithms seek to replicate. If human experts cannot agree on the weights, training an algorithm to match them becomes a nearly impossible task. Second, it can introduce bias into the job search process. Some countries, such as the United States, regularly require companies to prove the job search process is absent of any racial or gender bias. Some of the features identified, such as "no notable gaps in employment", "university attended", and "languages spoken", could easily introduce bias if left unchecked.

It may appear that selecting candidates is too nuanced for even the more sophisticated classification algorithms to perform well. However, algorithms can benefit both short- and long-term if humans are an integral part of the process. First, by filtering out the candidates who are clearly a mismatch for the advertised job position, the pool of candidates can be quickly restricted; thus, more attention can be put on evaluating those candidates that remain. Second, if the algorithm can learn which rules and patterns are absolute ("hard") and which can be have a probabilistic weight assigned ("soft") through human input, the algorithm can be trained to incorporate this feedback into the selection and ranking process. While algorithms such as learning to rank [16] can take candidate rankings produced by human experts and learn features automatically, the vast number of features relative to the size of the training set can lead to overfitting. Also, quickly filtering out candidates from the pool early on may provide too few samples for a classification algorithm to learn from.

Typically, the candidates who apply for a job position are the only ones considered. However, expanding the pool of candidates for the algorithm to consider to *all* applicants for a company (including those who applied for other positions) is one way to help the algorithm learn more quickly. Initially, it may seem this approach to expand the pool is unproductive; determining a short-list of candidates often involves evaluating each candidate *relative* to the other applicants in the pool (using features such as those indicated in Table 1). However, a larger set of applicants can increase the size of the training set by adding noise that is roughly Gaussian in nature, helping the algorithm to avoid overfitting [17]. A good training set should span the complete variability of each feature [18]; with a limited set of candidates, the rules the algorithm learns may become skewed, negatively affecting the ability for the algorithm to learn [19]. Obtaining a training set with candidates outside the pool of applicants can help the algorithm learn rules, even if those candidates are flagged for removal from final consideration.

Using a C4.5 classifier, we examine how this improvement can be made for the 5 job descriptions mentioned earlier. We wish to examine how the use of the set of applicants for *each* position and the set of applicants for *all* positions would improve our algorithm when humans were added to the loop. This approach attempts to answer our second research question.

3.1 Metrics

The list that most closely resembles the ranked list provided by our oracle is the best match. We use the RBO metric to determine a similarity score with the ranked list provided by our oracle. In addition, we have the experts provide a binary relevance (relevant/not relevant) for each job applicant. This allows us to also evaluate precision and recall and the F-measure, which is the harmonic mean of precision and recall.

3.2 Oracle

We asked a different set of 3 human HR experts collaborating as a team to rank the candidates for each job position. Rarely are more than 10 candidates brought in for interviews, so limiting the number of candidates to 10 is reasonable. These experts also evaluated each candidate's binary relevance (ether as relevant or not relevant) to the position, and the majority relevance label from the 3 experts is used as our oracle.

3.3 Baseline

For our baseline, we first eliminate candidates that don't meet the stated criteria for each position. For the candidates that remain, we randomly select a third of them and ask 3 human HR experts to independently rank and score relevance for each candidate for each of the 5 job positions. The two thirds not randomly selected are used as the test set. A C4.5 decision tree algorithm determines a final ranked set of candidates for each position and the binary relevance.

3.4 Treatments

Our first treatment (T1) is to use all the candidates who applied for *each* job description and randomly divide them into training and test sets with a ratio of 1:2. Rules are derived by the C4.5 algorithm based on training set. In addition, relevance is scored for each candidate. There is no human involvement.

Our second treatment (T2) is to use all the candidates who applied for *all* job description and randomly divide them into training and test sets with a ratio of 1:2. Rules are derived by the C4.5 algorithm based on training set. Once the test set is ranked, those who did not apply for the job are removed from the final candidate ranking. In addition, relevance is scored for each candidate. As with T1, there is no human involvement.

Our third and fourth treatments (T3 and T4) are similar to the first and second treatments, respectively, but they incorporate humans-in-the-loop inputs with the training sets by providing human feedback into the rules used in the C4.5 algorithm. Three human HR experts highlight the keywords and phrases that contributed to their ranking decision. The algorithm ranks the test set. In the case of T4, once the test set is ranked, those who did not apply for the job are removed from the final candidate ranking.

4 Results and Discussion

The overall RBO averages, and the precision, recall and F-measure for the baseline and 4 treatments are provided in Table 3.

Table 3. Average RBO score, precision, recall, and F-measure for the baseline condition and four treatments.

Condition/ Treatment	Average RBO Score	Average Precision	Average Recall	F-Measure
Baseline	0.305	0.550	0.808	0.654
Treatment 1	0.352	0.485	0.747	0.588
Treatment 2	0.373	0.495	0.790	0.609
Treatment 3	0.527	0.580	0.835	0.685
Treatment 4	0.576	0.600	0.857	0.706
Average	0.427	0.542	0.807	0.648

From Table 3, we see that treatment T4 provided the best average RBO scores, precision, recall, and F-measure scores. Treatment T2 provided better results than T1, particularly with RBO scores and recall, indicating the strength of having more data from which to train our algorithm. More impressively, RBO, precision and recall scores for treatments T3 and T4 improve upon treatments T1 and T2, illustrating the benefits of human-in-the-loop involvement in setting rules for our C4.5 classifier.

Comparing these treatments with our baseline in average RBO score, one can observe that the baseline falls in the middle of the pack; it outperforms treatments T1 and T2 but underperforms treatments T3 and T4. One can see that human-in-the-loop involvement to assist with establishing rules (as opposed to selecting candidates and letting the machine algorithm establish the rules from the selected candidates) can improve the rankings made by the machine algorithm. One also can observe that using a broader set of candidates for rule creation, then later eliminate those who did not apply for that position, can improve the rule set and subsequent candidate ranking. This supports the findings in [16], [20] and by other researchers exploring adding noise to classifiers in different contexts (e.g., [21]).

We note that using a single decision tree algorithm for different job positions can lead to poor classification decisions if the jobs descriptions have little in common with each other. In our case, all were for middle-management jobs supervising technical people; therefore, the variance in our RBO rankings between positions was a very reasonable 0.047.

5 Conclusion and Future Work

Our study examined the role of humans in selecting candidates for 5 middle-management job positions that relied on a combination of technical and soft skills. We used a C4.5 decision tree classification algorithm. The first part of our experiment examined how consistent are HR experts in determining the best features of a candidate's materials to evaluate. With respect to feature selection, we found little consistency through our group of 13 HR experts. This implies that one HR personnel could arrive at a very different list from another, bringing up concerns of potential bias and demonstrating the difficulties of coming up with an algorithm to completely replace humans. For the foreseeable future, machines will still need humans to be a part of the process.

The second part of our experiment examined how human experts might provide input to better train a machine and how the results might be further improved in job candidate selection able to improve the results. Our evaluation looked at similarity with separate machine/algorithm approaches with a set of human experts, finding when humans help establish the rules (as opposed to only selecting and ranking the relevant candidates), the ranked list, recall and precision scores improve. We also added other candidates who did not apply for the job position as noise (and removed these non-applicants at a later step), which helped improve the overall results and minimize overfitting. Thus, it is not only having humans in the loop, but having humans perform the most beneficial tasks, that best replicate our human experts.

In future work, we plan to examine the role of longitudinal information (the white box in Figure 1) as inputs to the decision process. Unfortunately, obtaining this information is challenging. Second, because most hiring is for non-management positions, we wish to see if our process can be replicated for blue-collar jobs as well. we wish to see better ways to test the effectiveness of our method's hiring suggestions; in other words, we wish to detect the best way to measure if the best person was recommended from the pool of candidates. One possibility is to examine internal hires in a longitudinal study (assuming data collection is possible), since in theory we can track the long-term career progression of the candidate within the firm offered the position as well as those that weren't. Third, we also plan to look at a wider variety of job positions and see how bias might potentially become part the algorithm. If we can detect bias early on, we can set some type of alarm to involve humans to correct for this. Fourth, we also wish to determine ways in which the algorithm can assist humans to do their job more effectively. One possibility is to explore how warnings can be provided for candidates that are either underqualified or overqualified based on some criteria established in advance. Another is to provide better graphical information for each candidate to illustrate their strengths and weaknesses relative to the candidate pool.

References

1. Stone, D. L., Deadrick, D. L., Lukaszewski, K. M., & Johnson, R. (2015). The influence of technology on the future of human resource management. Human Resource Management Review, 25(2), 216-231.
2. Zielinksi, D. (2017, February 13). Recruiting Gets Smart Thanks to Artificial Intelligence. https://www.shrm.org/resourcesandtools/hrtopics/technology/pages/recruiting-gets-smart-thanks-to-artificial-intelligence.aspx, last accessed 2018/07/24.
3. Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm. International Journal on Computer Science and Engineering, 2(08-2010), 2526-2534.
4. G. K. F. Tso and K. K. W. Yau. (2007) "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, pp. 1761-1768.
5. Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with applications, 34(1), 280-290.
6. The Economist. (2018, May 15). Special Report: AI in Business. https://coriniumintelligence.com/the-economist-special-report-ai-and-business/, last accessed 2018/07/24.
7. Azim, S., Gale, A., Lawlor-Wright, T., Kirkham, R., Khan, A., & Alam, M. (2010). The importance of soft skills in complex projects. International Journal of Managing Projects in Business, 3(3), 387-401.
8. Azzini, A., Galimberti, A., Marrara, S., and Ratti, E. (2018) A Classifier to Identify Soft Skills in a Researcher Textual Description. In: Sim K., Kaufmann P. (eds) Applications of Evolutionary Computation. EvoApplications 2018. Lecture Notes in Computer Science, vol 10784 Springer. DOI: 10.1007/978-3-319-77538-8_37
9. Wowczko, I. A. (2015). Skills and vacancy analysis with data mining techniques. In *Informatics* (Vol. 2, No. 4, pp. 31-49). Multidisciplinary Digital Publishing Institute.
10. Kelemenis, A., & Askounis, D. (2010). A new TOPSIS-based multi-criteria approach to personnel selection. Expert systems with applications, 37(7), 4999-5008.

11. M Kabak, M., Burmaoğlu, S., & Kazançoğlu, Y. (2012). A fuzzy hybrid MCDM approach for professional selection. Expert Systems with Applications, 39(3), 3516-3525.
12. Strohmeier, S., & Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research. Expert Systems with Applications, 40(7), 2410-2420.
13. Harris, C. (2011). You're hired! an examination of crowdsourcing incentive models in human resource tasks. In Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM) (pp. 15-18). Hong Kong, China.
14. Harris, C. G. (2017). Finding the Best Job Applicants for a Job Posting: A Comparison of Human Resources Search Strategies. In Data Mining Workshops (ICDMW), 2017 IEEE International Conference on (pp. 189-194). IEEE.
15. Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS), 28(4), 20.
16. Fuhr, N. (1992). Probabilistic models in information retrieval. The computer journal, 35(3), 243-255.
17. López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.
18. Choi, J., Rastegari, M., Farhadi, A., & Davis, L. S. (2013). Adding unlabeled samples to categories by learned attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 875-882).
19. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.
20. Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.