ECMLPKDD

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

7-11 September 2015, Porto, Portugal

Reviews For Paper

Track	PhD track
Paper ID	681
Title	Rankings of financial analysts as means to profits

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
What is you overall recommendation for this paper?	Reject
Provide a summary of the paper	In this work, the authors deal with the problem of ranking financial analysts. Instead of modeling the individual analyst's characteristics to predict the most accurate analysts, in the approach followed, the authoris characterize the genarl behavior of the rankings based on independent state variables. To perform ranking prediction, naive Bayes algorithm is used for label ranking having two functions, dealing with numerical attributes and with time series of label ranking data. In addition, the authors develop a trading strategy combining the predicted rankings with the BL model so as to form optimal portfolio. This strategy is applied to SP500 stocks.
List three strong points of the paper	1. The finance domain is of interest, especially to institutions and being able to effectively rank analysts has a good potential for the investors. 2. The idea of using independent state variables to characterize the general behavior of the rankings of analysts, instead of paying attention just to the individual anaslyst's characteristics, is useful.
List three weak points of the paper	1. Notation: the symbols and notation used throughout the paper makes it very hard to read and creates continuously confusion 2. Content: there are several gaps in the explanation of concepts and many details are missing, making the paper at times hard to understand and follow. The flow should be also definitely improved. Examples would make the paper better. Please see the detailed comments. 3. Experimental Results: there is no particular explanation/interpretation/discussion of the results. This makes the paper even weaker.
	Content 1. Introduction A reference should be given for SP500 (it's the 1st time that it's mentioned). 3. Label ranking alrgorithm - "x_t = {...} instances" --> either remove instances or rephrase. x_t refers

to the vector of variables so that should be clear

- The independent variables are referred to as x_1...x_4. However, prior to that the instance space of variables is denoted as {V_1,...V_m}, and x_i are the examples in the training set. The variable is mentioned as \alpha. The same holds for Table 1, where the variables are referred to as x_1, ..., x_4. Apart from that, based on the {V_1,...,V_m}, one would expect the variables to be 1,...,m, or, at least, since \alpha is used, to make somehow \alpha range from 1 to m. There is a confusion here on all of the above and changes are needed.

- Using x_i in the equations (from (1) to (6)) makes things more complicated. My understanding is that the authors want to refer to a vector of variable values and based on that to predict y. However, x_i was used earlier in the text to represent a training example. This should be avoided, and it's better to use just x.

- In Equation (2), the function for measuring the similarity between rankings should be explicitly given (both the notation, before being used in (2), and an equation of how it's computed).

- Just before Equation 3, the following should be changed. Again, the authors refer to variable x_i, while earlier they refer to variable \alpha for V_\alpha. In addition, the attribute \alpha is used. If the attribute is considered to be the same as a variable, then that should explicitly described. Moreover, the (v_{i,a}) corresponds the value for x_{i,\alpha}. However, in the previous page the domain of a nominal variable \alpha is given as v_{\alpha,1}, ..., v_{\alpha,n_\alpha}. This comes in contrast to (v_{i,a}), i.e., opposite subscripts used, while using i may not be (as mentioned before) the best option. x_i is an example, but earlier it was referred to as an training example that implies that we do know its output y_i and we don't have to predict it.

- In Equation (4), m is used for subscript, but m is the domain space of the last variable m, according to the notation introduced in the beginning of section 3. The same holds for equations (8) and (9). Thus, a more "general" letter should be used, since the authors want to refer to any variable.

- In Equation (7) the authors replace essentially y of Equation (2) with y_t. If that's the case, then what is the y used in the function measuring the similarity between rankings? This part is also confusing with notations. Also, t ranges from 1 to n, i.e., represents the examples. If so, if the authors want to use t for time, then using the same variable for enumerating the examples makes things more complicated.

- By referring to n examples in Equation (7), do the authors refer to all the training examples that have come up to t (up to time t-1)? And, if so, is n-the example the latest one and 1 is the oldest one?

- w_t = {w_1, ..., w_n} -> w is not correct, and in particular the use of w_t. In Equation (7), w_t is used to express weight for example t, based on my understanding. So, using the same notation to represent a set of weights is not correct.

- In the definition of vector w, the notation {n}_1^n is not explained. Is it a set, or a numeric value representing the training example? If it represents an example, should that be {t}_1^n? And, again, the use of t is not the best possible one.

- It is mentioned that the larger b the more weight is given to more recent rankings. If the most recent ranking is considered to be numbered as n, then, based on the equation for computing the vector w, doesn't seem to represent

Provide detailed comments to the authors	that. The exponent has the {variable}, which increases (assuming that 1 is the oldest example and n the latest) and gives larger negative values in the whole exponent. As a result the value of b becomes smaller for the latest rankings. Maybe I'm missing something here, but definitely more explanation should be given in this part, and, again, better notation should be used. - In Equation (9), there is the "sigma", while it should be "\sigma". 4. Trading Strategy - In step 1, it is mentioned that the rankings of all analysts is predicted. But how exactly is that done? There is no connection between the previous section and this one. More details should be given. How and which training examples are used? What are the variables and their values so as to predict y? Are the values of the variables in the beginning of the quarter used? These are important missing details. - In step 2, the authors state taht these predicted rankings and analysts' price targets are used to define essentially the views. But how exactly are the expected returns computed? - The notation s is used in the Q_{q,s} is used, but s is not explained. It's the stock, but that should not be left to the understanding of the reader. 5. Data and experimental setup - How exactly is the "Target rankings" used (section 5.1)? - In the notation used for the actual price t is inserted in Equation (11) but - After Equation (13) more details should be given on the ranking happening (that these values are ordered for all brokers and which broker gets the highest/worst ranking). Such details would make the paper more complete. - Section 5.2 should be presented earlier in the paper (before the trading strategy). - Are all variables mentioned used in the experiments? I would expect more details on the variables, their range of values, or an example. - Why are the methods for the dynamics of the variables are referred to as "aggregation methods"? - Again, x_t is used for the raw method to refer to a variable. - More details should be given on how these 4 methods are used for the dynamics of the independent variables. Examples would be useful to have. - Are the information sets the training examples used? If so, that should be made more explicit. 6. Empirical Results - Why is the section named "Empirical Results" and not "Experimental Results"? - table (2) --> Table 2 - No information is given on "turnover ratio". - In the analysis, a few numbers are given but no actual and meaningful

	interpretation is given comparing the methods. Although some results are obvious, more discussion should be given as text. - In Fig. 1 you can use shapes, and not just lines for the methods, so that the figure is readable in black and white as well. 7. Conclusion - Refer earlier in the text to the "baseline ranking prediction" (last sentence). English Abstract - "individual analyst'" --> "individual analyst's" 1. Introduction - p.2: "to predict the the rankings" --> remove "the" 3. Label ranking algorithm - p.3: "ranking of k label at time t" --> "k labels" - p.4: "there has been established" --> "a connection has been established" - p.4: avoid having "in (1)"; rather rephrase to "in Equation (1)" 4. Trading Strategy - "risk-free rate for the same" should be made more formal 5. Data and experimental setup - "until it revised or" --> "until it is revised" - In Section 5.3, "about the analyst' performance" --> "about the analyst's performance" 6. Empirical Results - "Observe that" --> make it more formal - "of the strategy achieved for the method" --> "was achieved" References - In section 3.1, "Following [1]" should be replaced with the names of the authors before the citation, or put [1] at the end of the sentence
If accepted, this work should appear as:	Poster

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
What is you overall recommendation for this paper?	Accept
Provide a	The manuscript studies the problem of making money on the stock market. More particularly, the authors study how to augment the ranking of financial analysts with time series of past performances. The authors propose to deal with the associated learning problem by using the Naive Bayes algorithm for Label Ranking. In order to deal with the continuous values they assume a

summary of the paper	normal distribution and the time series are dealt with by weighting the similarity scores at different time points. They then integrate the predictions with a portfolio management model in order to find a good trade-off between risk and projected profit. The results are analysed in-depth, including analysis of the impact of some variable modelling decisions, and the results on past data are clearly much above market performance.
List three strong points of the paper	- The topic is quite novel as an application in the conference. - The paper is very well written. - The approach is quite simple but it is perhaps unexpected that such a simple approach could lead to such benefits.
List three weak points of the paper	- The approach is a bit simple and ad-hoc (little fundamental motivation).
Provide detailed comments to the authors	I like the paper. It would be nice if it would be presented at the PhD forum. Regarding the conversion from Euclidean distance to probability; I would expect it is broadly known in the ML community that the normal distribution has a squared difference in the exponent and can hence model squared differences, if the variance is fixed and the variables are independent; I think this is known as an isotropic Gaussian. I was surprised to find a reference to a chemistry paper in this context (Section 3.1) and the corresponding claim that there exists a `connection between probabilities and the general Euclidean distance' is a bit too strong in my opinion. If the authors believe that their method really works and will continue to be generally applicable in the future, I would definitely suggest them to bring a tool to the market as soon as possible and patent the technology as much as possible.
If accepted, this work should appear as:	Regular Paper