# Learning of Performance Measures from Crowd-sourced Data with Application to Ranking of Investments

Greg Harris[1], Anand Panangadan[2], and Viktor K. Prasanna[2]

[1] Department of Computer Science
University of Southern California, Los Angeles, California
gfharris@usc.edu
[2] Ming-Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, California
{anandvp,prasanna}@usc.edu

**Abstract.** Interestingness measures stand as proxy for "real human interest," but their effectiveness is rarely studied empirically due to the difficulty of obtaining ground-truth data. We propose a method based on learning-to-rank algorithms that enables pairwise rankings collected from domain community members to be used to learn a domain-specific measure. We apply this method to study the interestingness measures in finance, specifically, investment performance evaluation measures. More than 100 such measures have been proposed with no way of knowing which most closely matches the preferences of domain users. We use crowd-sourcing to collect gold-standard truth from traders and quantitative analysts in the form of pairwise rankings of equity graphs. With these rankings, we evaluate the accuracy with which each measure predicts the user-preferred equity graph. We then learn a new investment performance measure which has higher test accuracy than the currently proposed measures, in particular the commonly used Sharpe ratio.

## 1 Introduction

The goal of data mining is to automatically identify "interesting" patterns in a dataset. Data mining algorithms therefore utilize an *interestingness measure*, a function that assigns a numerical score to a given pattern, to evaluate and rank patterns. Several interestingness measures have been proposed, surveyed, and evaluated for different domains [16, 27, 20, 13, 19, 26, 3]. The choice of interestingness measure depends on the specific domain since a pattern can exhibit multiple desirable attributes which must be traded-off against each other.

Designing an interestingness measure for a specific domain is challenging and typically requires a domain expert to create a new function and identify a set of features that can be calculated from the dataset attributes [22]. As an alternate approach, we propose a method to *learn* an interestingness measure from crowd-sourced data collected from end-users in the domain community. In our approach, domain users are presented with pairs of candidate patterns and

are asked to rank one over the other. Pairwise ranking is a non-arduous way for domain users to share preference information. It also facilitates the combining of preference information from multiple users. The collected pairwise rankings are then provided as input to a learning-to-rank algorithm to learn a model of user preference which can be used as an interestingness measure. The features in the learning model are previously proposed interestingness measures for the domain. The result is a custom measure that represents "real human interest" [22] in the domain as expressed by its users.

We demonstrate the proposed approach and evaluate its effectiveness in the domain of finance, specifically the task of learning an investment performance measure that reflects the preferences of investment professionals. Investment preference rankings are collected from users of online discussion forums comprised of quantitative analysts and traders. The model features that are used in the learning-to-rank algorithm include currently used investment performance metrics and ratios. The learned model achieves an accuracy of 80% for predicting the domain users' preference, while the highest accuracy of any single existing performance measure is 77%.

We believe that learning such an interestingness measure can benefit this domain since there is a large number of investment choices. For instance, the United States has over 5,000 exchange-traded stocks and over 7,000 mutual fund choices. Our proposed approach can enable individuals to locate investments that match their specific interests. Moreover, the learned interestingness measure can also be used as an objective function for portfolio selection and optimization.

The contributions of this work are as follows:

1. We propose a novel approach based on learning-to-rank algorithms that enables a domain-specific performance measure to be learned from domain community contributions. The method requires only pairwise preferences from domain experts.
2. We evaluate this approach in the domain of investment ranking and show that the learned performance measure has higher accuracy than existing domain-specific measures. We also address issues of data quality that are critical in crowd-sourced datasets.
3. We provide all data collected as part of this study to encourage further research in this area[3].

## 2 Related Work

Ohsaki et al. [22] experimentally compared interestingness measures against *real human interest* in medical data mining. They generated prognosis-prediction rules from a clinical dataset on hepatitis. They then had a medical expert evaluate rules as *Especially-Interesting*, *Interesting*, *Not-Understandable*, and *Not-Interesting*. Carvalho et al. [5] build on [22] with evaluations on eight datasets. They presented nine rules to each expert for each interestingness measure: the

[3] `http://thames.usc.edu/rank.zip`

best three, the worst three, and three in the middle. Experts were asked to assign a subjective degree of interestingness to each rule. Tan et al. [26] studied ways to select the best interestingness measure for association rules – instead of using actual experts to rank contingency tables, they consider a held-out measure as the expert (and repeat over all measures). None of these works attempt to *learn* an interestingness measure from domain experts as we propose in this work.

To the best of our knowledge, no work has been published on comparing investment performance measure rankings against real human interest. For related work in finance, we summarize publications that describe the relative performance of different evaluation measures in this domain. Justification for these proposed measures is axiomatic, based on the properties of the measures [1, 17]. Farinelli et al. [11] compare eleven performance ratios. Their work includes a limited empirical simulation, evaluating how well each ratio performed forecasting five stock indexes. They find that asymmetrical performance ratios work better and recommend that more than a single performance ratio be used. Cogneau and Hübner [7] survey over 100 investment performance measures. They provide a taxonomy and classification of measures based on their objectives, properties, and degree of generalization. Bacon [2] also provides a thorough survey of measures grouped into categories.

Some of the current research indicates that different performance metrics produce substantially the same rank orders. Hahn et al. [14] used 10 performance measures to rank data from two proprietary trading books and found high values of Spearman's rank correlation. Eling and Schuhmacher [10] find high rank correlation (0.96) between 13 performance measures that were used to rank the returns of 2,763 hedge funds. Eling [9] confirmed the high rank correlation between measures when applied to 38,954 mutual funds from 7 asset classes. On the other hand, Zakamouline [28] describe several less correlated measures and suggest the use of Kendall's tau instead of Spearman's rho for measuring rank correlation. None of these four studies considered the Pain, Ulcer, and Martin-related measures discussed in Section 3.3.

## 3 Finance Background

Investment performance measures are designed to weigh the risk as well as the reward, and are therefore called "risk-adjusted returns." Metrics are structured as ratios, with return on investment in the numerator, and risk in the denominator. In this way, a single metric can compare two investment options with different risk profiles.

While return on investment is a standard measure of reward, there are multiple measures of risk and hence consensus has not yet been reached as to which performance measure is best [11]. New performance metrics continue to be proposed [7, 21], and investors have to choose from among them [2].

We first describe equity graphs which provide a visualization of asset performance, followed by a summary of performance measures that will be used as features in our learning model.

### 3.1 Equity Graphs

Historical performance is often presented as an *equity graph*, which shows the value of one's investment account over time. Equity graphs enable domain experts to rapidly evaluate historical performance. While there are different types of equity graphs, in our work we use the common variant where the graph presents a cumulative sum of daily returns. This is equivalent to assuming exactly one dollar was invested each day, with profits removed from the account. Such a graph is easy to examine, since the ideal is a straight line from the lower left corner to the upper right corner. Examples are shown in Figures 1, 2, and 3.

### 3.2 Distribution-Based Measures

Many performance measures calculate risk based on the distribution of *returns*. For a time series $R$, the return on investment for each period, $R_t$ is:

$$R_t \equiv \frac{S_t - S_{t-1}}{S_{t-1}}$$

where $S_t$ is the asset value at time $t$.

The baseline investment performance measure is the reward to variability ratio, the *Sharpe ratio* [23]. The Sharpe ratio is widely used [9], with surveys showing its use by up to 93% of money managers [2]. This performance measure is "optimal" if the return distribution is normal. The Sharpe ratio is closely related to the $t$-statistic for measuring the statistical significance of the mean differential return [24].

Using the same notation as Sharpe [24], let $R_{Ft}$ be the return of the investment in period $t$, $R_{Bt}$ the return of the benchmark security (commonly the risk-free interest rate) in period $t$, and $D_t$ the differential return in period $t$:

$$D_t \equiv R_{Ft} - R_{Bt}$$

Let $\bar{D}$ be the average value of $D_t$ from period $t = 1$ through $T$:

$$\bar{D} \equiv \frac{1}{T} \sum_{t=1}^{T} D_t$$

and $\sigma_D$ be the standard deviation over the period:

$$\sigma_D \equiv \sqrt{\frac{\sum_{t=1}^{T} (D_t - \bar{D})^2}{T-1}}$$

The Sharpe Ratio ($S_h$) is:

$$S_h \equiv \frac{\bar{D}}{\sigma_D}$$

Many performance evaluation measures are modifications of the Sharpe ratio. Given that asset returns are often non-normal, researchers have developed

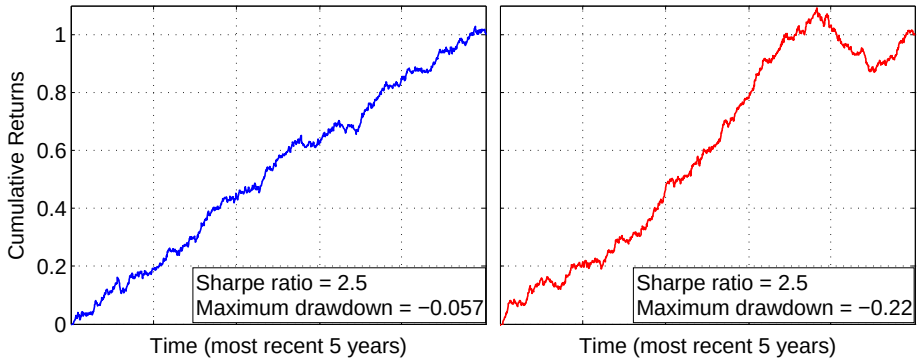


Fig. 1: The red chart on the right was generated by randomly permuting the daily returns from the blue chart on the left. Both have the same distribution of daily returns, and hence the same daily Sharpe ratio. This figure illustrates how distribution-based performance measures cannot capture some features preferred by traders, such as a small maximum drawdown.

measures that incorporate higher moments of the distribution [17]. The Sortino ratio [25] is similar to the Sharpe ratio, except it uses the semi-standard deviation (downside risk) in the denominator. Other measures consider only the very worst returns in the tail of the return distribution [8, 1].

### 3.3 Multi-Period-Based Measures

Shape-based measures focus on multi-period drawdowns instead of return distributions. The *Maximum Drawdown* is defined as the maximum peak-to-valley decline in the equity graph. Figure 1 shows how two orderings of returns can have very different maximum drawdowns while still having the same daily Sharpe ratio. The chart on the right has an unappealing drawdown of 22%, yet it has the exact same distribution of returns as the chart on the left (with a drawdown of only 6%).

*Drawdown* can also be defined as a string of consecutive negative returns. Many performance measures consider aspects of the distribution of such drawdowns instead of returns, including the mean, standard deviation, and selected number of worst drawdowns.

The *Martin ratio*, or "Ulcer performance index" has the same numerator as the Sharpe ratio, but has the *Ulcer index* as the denominator. Using the notation in Bacon [2], let $D_i'$ be the drawdown since the previous peak in period $i$. The Ulcer index is then defined as:

$$\text{Ulcer index } UI = \sqrt{\sum_{i=1}^{n} \frac{D_i'^2}{n}}$$

Figure 2 shows an equity graph with each $D'_i$ shown in black. The Ulcer index penalizes long drawdowns.

The *Pain ratio* also has the same numerator as the Sharpe ratio. The denominator is the *Pain index*, a modified form of the Ulcer index:

$$\text{Pain index } PI = \sum_{i=1}^{n} \frac{|D'_i|}{n}$$

The Pain index also penalizes long drawdowns but does not penalize deep drawdowns as severely as the Ulcer index.

*Max Days Since First at This Level* is an intuitive measure that we define as the longest horizontal line that can be drawn between two points on the graph, as shown in Figure 3. We introduce it here because it is not found in the literature, and we find it ranks highly in our experiments.

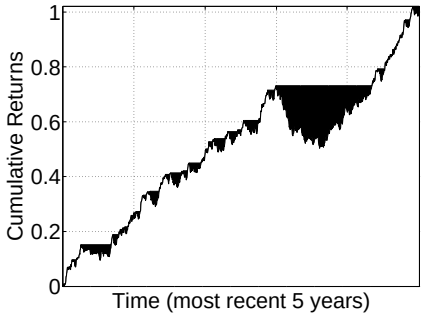


Fig. 2: The Pain index is the area colored black. The Ulcer index is the root mean squared height of each vertical black line.

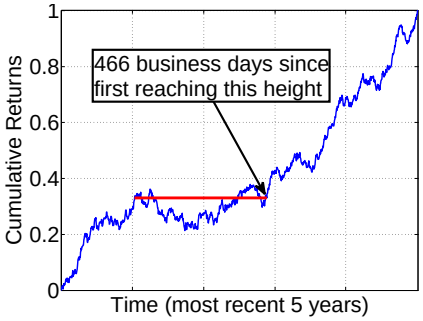


Fig. 3: "Max Days Since First at This Level" is the longest horizontal line that can be drawn between two points on the graph.

## 4 Approach

We now describe our approach to learn an investment performance measure with higher rank prediction accuracy than the current performance measures, using crowd-sourced domain user input. The steps of our approach are as follows:

1. Generate equity graphs simulating reasonable investment performance.
2. Collect preference data for the generated equity graphs from domain users in the form of pairwise rankings.
3. Use learning-to-rank algorithms with individual performance measures as features to create a new performance measure.

### 4.1 Generating Equity Graphs

Our approach uses equity graphs as a means for enabling domain experts to rapidly compare two strategies or investments. We generated (synthetic) equity graphs that follow a log-normal random walk. In this model, the asset price, $S_t$, follows the stochastic differential equation:

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

where $\mu$ is the constant drift, $\sigma$ is the constant volatility, and $dW_t$ is a Wiener process.

We generated discrete differential simple returns representing five years with 252 business days per year. The returns are normally distributed with a mean of 0.125 and a standard deviation of 1. These values were chosen to lead to a broad distribution of Sharpe ratios centered around 2. Of these, only graphs with Sharpe ratios between 1.5 and 2.5 are retained. This range corresponds to the range of Sharpe ratios typically encountered. Ratios below 1.5 are unattractive as an investment, and ratios greater than 2.5 are very rare in practice. In total, we generated 2,000 charts.

For each graph, we normalize the set of returns to sum to 1. Normalizing the cumulative return enables domain experts to directly compare risk metrics (such as the maximum drawdown) on the same scale.

### 4.2 Collection of Ranking Data

One of our innovations is the collection of domain expert preferences in the form of pairwise rankings. We believe that it is easier for a participant to choose between two equity graphs than to decide on a numeric score for every individual graph. In particular, numeric scores require that these be normalized before aggregating scores to account for the different preference scales of participants. This normalization would be difficult for cases where a participant only labeled a small number of charts. In contrast, our pairwise ranking-based method is fast for human users with median ranking time between 3 and 4 seconds.

We created a web page that described our research goal and presented two randomly chosen equity graphs side-by-side. A participant is asked which of these two investments is more attractive to invest in for the future. We requested participation from domain experts in two online forums. The first forum targets quantitative analysts and risk managers. The second forum targets individual traders, although some members run small hedge funds or are commodity trading advisors. 66 different anonymous people from these forums ranked a total of 1,004 chart pairs. We believe that the participation of many professionals is validation of community interest in improving investment performance measurement.

One author also ranked 1,659 equity graph pairs, including a re-ranking of every pair ranked by the community. In order to estimate self-consistency of rankings, the author later re-ranked each of the same 1,659 graph pairs. The estimate of self-consistency is 90%. In all rankings and re-rankings, the equity graph positions (i.e., left or right side) were chosen randomly.

### 4.3 Data Quality

Ensuring quality of crowd-sourced data is a recognized problem [18]. As expected, we found that some of the crowd-sourced data was of low quality. In this section, we describe the steps performed to derive a higher quality data subset from the crowd-sourced annotations.

One author tagged each of the pairs of equity graphs used for crowd-sourced ranking as either "close call" (81%) or "clear choice" (19%). A "clear choice" tag indicates that the author's preference was strong and this view was likely to reflect universal preferences. The author was 100% self-consistent when re-ranking "clear choice" equity graph pairs.

To identify low quality contributions, we evaluated each contribution according to the following characteristics:

- Small median time between clicks
- A high fraction of times the participant clicked the same button (i.e., left or right), rather than alternating approximately uniformly between the two
- A systematic preference for the chart with the lower Sharpe ratio
- A relatively high fraction of rankings that contradict the author's "clear choice" rankings

Overall, we filtered out 129 rankings, leaving 875 of the original 1,004. As such a data quality filter is subjective, we also ran all experiments on the unfiltered dataset in addition to making the data publicly available.

### 4.4 Learning-to-Rank

A *learning-to-rank* algorithm predicts the order of two objects given training data consisting of partial orders of objects (and their features). We use the learning-to-rank algorithm proposed by Herbrich et al. [15]. In this method, the ranking task is transformed into a supervised binary classification task by considering the *difference* between corresponding features. This transformation also enables the use of other learning algorithms in addition to support vector machines as originally proposed by Herbrich et al. [15].

The three classification algorithms we use in this work are:

1. Logistic regression, with $L_1$-norm regularization [12]
2. Random forests [4]
3. SVM with linear and RBF kernels [6]

Given two objects, $A$ and $B$, the learning-to-rank task is to predict if $A > B$ based on their respective features. It is redundant to include both $A > B$ and $B > A$ (with negated feature differences) when training a model. In order to ensure balanced numbers of classes for the model to learn, we chose one of either $A > B$ or $B > A$ for each instance such that there were equal numbers of positive and negative instances in the training data. Balancing the training data also ensures that the intercept or bias term will be zero for logistic regression.

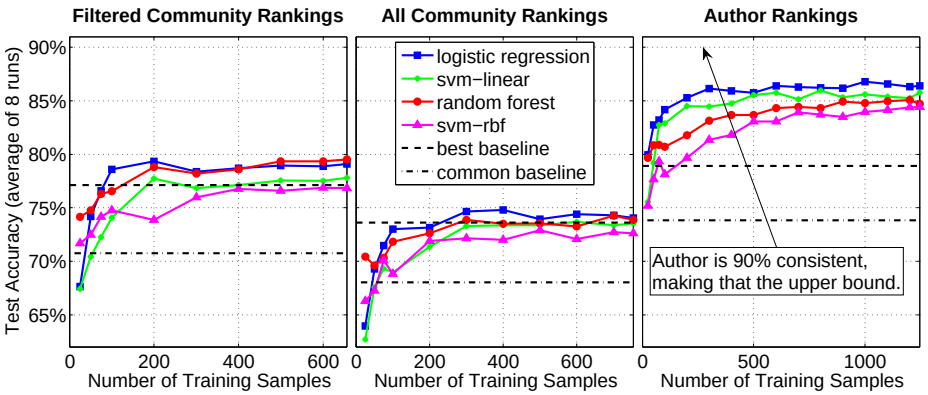


Fig. 4: Accuracy of learning-to-rank models trained and tested on crowd-sourced "real human interest" data in the form of pairwise rankings.

**Features** The features we use as inputs to the machine learning models include relevant risk and performance metrics found in Bacon's comprehensive survey [2] which also provides descriptions of each measure using uniform notation. Note that for our normalized charts, risk metrics produce identical rank orderings as their respective performance measures. We nevertheless include both, because models such as logistic regression use linear combinations of features, and we do not know *a priori* which feature will combine best with other features.

## 5 Experiments and Results

In our experiments, we consider the following three datasets:

1. The full set of all 1,004 community rankings (ACR)
2. The filtered set of 875 community rankings (FCR)
3. The set of 1,659 author rankings (AR)

Each experiment followed these steps for evaluation:

1. Randomly shuffle the data
2. Separate 25% of the data for testing
3. Choose optimal hyper-parameters using 5-fold cross-validation on the training data
4. Test the accuracy of the final model on the held-out test data

We performed each experiment 8 times and averaged the test accuracies. All models were trained and tested on the same random shuffle of the data to better compare their accuracies.

In order to estimate the impact of the number of pairwise rankings needed for training on the accuracy of the learned performance measure, we tested progressively increasing amounts of training data. The data was not reshuffled

as training instances were added, i.e., for $n = 200$, the first 100 data points are the same ones used for $n = 100$. Figure 4 shows accuracies obtained for each of the three datasets, using each of the models, trained with an increasing number of pairwise ranking samples. Each point on the graphs represents the average of 8 runs. For reference, we show the most commonly used performance measure as a baseline, the monthly Sharpe ratio. In addition, we also show the performance of the *ex post facto* best measure for each dataset, although in practice which measure would perform the best on a given dataset would not be known.

From these experiments, we observed that none of the established performance measures in this domain is able to fully predict domain expert preferences. Our performance measure trained from domain expert preferences is able to achieve better prediction accuracy. For the filtered community ranking dataset, the random forests approach narrowly outperformed logistic regression, with 80% accuracy. The best baseline for this dataset is the monthly Pain index, with 77% accuracy. For the dataset containing all community rankings, logistic regression has the best performance, with 74% accuracy. The best baseline for this dataset is the daily Pain index, with 74% accuracy. For the dataset containing author rankings, logistic regression again has the best performance, with 86% accuracy. Note that for this dataset, the same author performed each pairwise ranking twice. As these two sets of rankings have an agreement rate of 90%, this forms an upper bound for any model's predictive accuracy. The best baseline for this dataset is the daily Martin ratio, with 79% accuracy.

Learning-to-rank accuracies are lower for the community datasets than the author dataset. This is because community members have idiosyncratic preferences, contributing inconsistency to the community training and test data.

The learning curves in Figure 4 are relatively flat. This indicates that ranking more equity graph pairs would not lead to higher accuracies, given the models and features we have chosen. A small number of rankings (approximately 300) is adequate to learn a trader's preferences. Given median ranking times between 3 and 4 seconds, a trader would likely spend 15 to 20 minutes ranking 300 chart pairs.

## 6 Conclusion

We presented a novel method using crowd-sourcing to learn a domain-specific performance measure. This method uses pairwise learning-to-rank algorithms with previously proposed performance measures as input features. We demonstrated and evaluated this approach for the case of learning a performance measure to rank investments. Our experimental results showed that machine learning algorithms can find linear combinations of performance measures that improve accuracy in this domain.

We provide all data[4] (equity graphs, measure calculations, and rankings) to encourage further study. With the data, we also include a table unable to

[4] `http://thames.usc.edu/rank.zip`

fit in this paper, showing the accuracy of the individual baseline performance measures on each dataset.

## Acknowledgment

We acknowledge the help of 66 different anonymous people from two online forums who provided the training data.

This work is supported by Chevron USA, Inc. under the joint project Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

## References

1. Gordon J Alexander and Alexandre M Baptista. Portfolio performance evaluation using value at risk. *The Journal of Portfolio Management*, 29(4):93–102, 2003.
2. Carl R Bacon. *Practical Risk-adjusted Performance Measurement*. John Wiley & Sons, 2012.
3. Julien Blanchard, Fabrice Guillet, Regis Gras, and Henri Briand. Using information-theoretic measures to assess association rule interestingness. In *Data Mining, Fifth IEEE International Conference on*, page 8. IEEE, 2005.
4. Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
5. Deborah R Carvalho, Alex A Freitas, and Nelson Ebecken. Evaluating the correlation between objective rule interestingness measures and real human interest. In *Knowledge Discovery in Databases: PKDD 2005*, pages 453–461. Springer, 2005.
6. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
7. Philippe Cogneau and Georges Hübner. The (more than) 100 ways to measure portfolio performance. part 1: standardized risk-adjusted measures. *Journal of Performance Measurement*, 13(Summer), 2009.
8. Kevin Dowd. *Beyond Value at Risk: The New Science of Risk Management*, volume 3. Wiley Chichester, 1998.
9. Martin Eling. Does the measure matter in the mutual fund industry? *Financial Analysts Journal*, pages 54–66, 2008.
10. Martin Eling and Frank Schuhmacher. Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking & Finance*, 31(9):2632–2647, 2007.
11. Simone Farinelli, Manuel Ferreira, Damiano Rossello, Markus Thoeny, and Luisa Tibiletti. Beyond Sharpe ratio: Optimal asset allocation using different performance ratios. *Journal of Banking & Finance*, 32(10):2057–2063, 2008.
12. Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.
13. Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.
14. Carsten Hahn, F Peter Wagner, and Andreas Pfingsten. An empirical investigation of the rank correlation between different risk measures. In *EFA 2002 Berlin Meetings Presented Paper*, pages 02–01, 2002.

15. Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems*, pages 115–132, 1999.
16. Robert J Hilderman and Howard J Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In *Advances in Knowledge Discovery and Data Mining*, pages 247–259. Springer, 2001.
17. Con Keating and William F Shadwick. A universal performance measure. *Journal of Performance Measurement*, 6(3):59–84, 2002.
18. Matthew Lease. On quality control and machine learning in crowdsourcing. In *Human Computation*, 2011.
19. Philippe Lenca, Patrick Meyer, Benoit Vaillant, and Stéphane Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610–626, 2008.
20. Kenneth McGarry. A survey of interestingness measures for knowledge discovery. *Knowledge Eng. Review*, 20(1):39–61, 2005.
21. Janki Mistry and Jubin Shah. Dealing with the limitations of the Sharpe ratio for portfolio evaluation. *Journal of Commerce and Accounting Research*, 2(3):10–18, 2013.
22. Miho Ohsaki, Shinya Kitaguchi, Kazuya Okamoto, Hideto Yokoi, and Takahira Yamaguchi. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In *Knowledge discovery in databases: PKDD 2004*, pages 362–373. Springer, 2004.
23. William F Sharpe. Mutual fund performance. *Journal of Business*, pages 119–138, 1966.
24. William F Sharpe. The Sharpe ratio. *Journal of Portfolio Management*, (21):49–58, 1994.
25. Frank A Sortino and Robert Van Der Meer. Downside risk. *The Journal of Portfolio Management*, 17(4):27–31, 1991.
26. Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 32–41. ACM, 2002.
27. Benoît Vaillant, Philippe Lenca, and Stéphane Lallich. A clustering of interestingness measures. In *Discovery Science*, pages 290–297. Springer, 2004.
28. Valeri Zakamouline. The choice of performance measure does influence the evaluation of hedge funds. *Available at SSRN 1403246*, 2010.