Ertek, G., Kaya, M., Kefeli, C., Onur, Ö., Uzer, K. (2012) "Scoring and predicting risk preferences" in Behavior Computing: Modeling, Analysis, Mining and Decision. Eds: Longbing Cao, Philip S. Yu. Springer.

Scoring and Predicting Risk Preferences

Gürdal Ertek[1], Murat Kaya[1], Cemre Kefeli[1], Özge Onur[1], and Kerem Uzer[2]

[1]Sabancı University, Faculty of Engineering and Natural Sciences,

Orhanli, Tuzla, 34956, Istanbul, Turkey.

[2]Sabancı University, School of Management,

Orhanli, Tuzla, 34956, Istanbul, Turkey.abanci University

Scoring and Predicting Risk Preferences

Gürdal Ertek[1], Murat Kaya[1], Cemre Kefeli[1], Özge Onur[1], and Kerem Uzer[2]

[1] Sabancı University, Faculty of Engineering and Natural Sciences,
Orhanli, Tuzla, 34956, Istanbul, Turkey. ertekg@sabanciuniv.edu
[2] Sabancı University, School of Management,
Orhanli, Tuzla, 34956, Istanbul, Turkey.

Abstract. This study presents a methodology to determine risk scores of individuals, for a given financial risk preference survey. To this end, we use a regression-based iterative algorithm to determine the weights for survey questions in the scoring process. Next, we generate classification models to classify individuals into risk-averse and risk-seeking categories, using a subset of survey questions. We illustrate the methodology through a sample survey with 656 respondents. We find that the demographic (indirect) questions can be almost as successful as risk-related (direct) questions in predicting risk preference classes of respondents. Using a decision-tree based classification model, we discuss how one can generate actionable business rules based on the findings.

1 Introduction

Financial institutions such as banks, investment funds and insurance companies have been using surveys to elicit risk preferences of their customers[3]. They analyze the collected data to categorize their customer pool and to offer customized financial services. For instance, the institution can emphasize safety and predictability of investments for customers who are categorized as risk-averse, whereas it can emphasize potential gains to customers who are categorized as risk-seeking. Determining customers' risk preferences is a prerequisite for developing healthy financial plans. For this purpose, leading financial institutions often integrate the survey results into their Customer Relations Management (CRM) systems.

While the use of financial risk preference surveys is popular in practice, the survey questions are rarely determined using scientific reasoning. In addition, when *risk scores* are calculated for survey respondents, questions are often given identical weights. Evaluating 14 risk surveys in France, [26] determines that *"Only a minority of the questionnaires in our sample rely on scoring techniques that attribute points for each answer. Furthermore, the questionnaires under review that do rely on scoring techniques generally fail to use sufficiently sophisticated econometric methods when setting their scoring rules. . . . Consequently, the classification of investors is still based on subjective judgments, rather than on data and quantified findings. "*. [26] also finds weak correlation between the risk scores of different surveys. That is, different financial institutions might be providing different financial advice to the same individual.

[3] See, for example http://www.paragonwealth.com/risk_tolerance.php

These observations indicate the need for scientific quantitative approaches for calculating risk scores using survey data. In this research, we offer a methodology to determine weights for the questions of a given risk survey, applying a regression-based iterative algorithm. Using these weights, we calculate a risk score for each survey respondent, which can be used for classification purposes.

Risk preference surveys include questions on two sets of respondent attributes: (1) *Direct attributes*, such as a choice between different hypothetical investment options, that are directly related to risk preferences; (2) *Indirect attributes*, such as demographic information, that are not directly related to risk preferences. The questions on direct attributes presumably provide more valuable information on respondents' risk preferences. However, since these questions aim at sensitive information and involve hypothetical scenarios, it may be difficult to elicit truthful information from respondents. This is particularly the case when the questions are numerous and framed too broadly. In contrast, indirect data is often readily available or can be collected easily. Our research offers a method to classify individuals based on their answers to indirect questions. One can use this classification to ask more tailored direct questions, if necessary.

The definition of risk, and risk preferences is context-dependent. Risk can be defined in many ways, including expected loss, expected disutility, probability of an adverse outcome, combination of events/consequences and associated uncertainties, uncertainty of outcome of actions and events, or a situation or event where something of a human value is at stake and the outcome is uncertain [3]. In this study, we focus on risk preferences of individuals in the context of financial investments.

The contributions of this work can be summarized as follows:

- We develop a novel behavior computing [5] methodology for scoring and prediction of risk preferences. The two main components of the methodology are:
  - *Risk scoring algorithm*: Given a risk survey, this iterative algorithm determines which questions (attributes) to use and the weights for each direct question, and calculates risk scores for all respondents based on these weights.
  - *Classification model*: This model classifies respondents based on a set of (direct, indirect or both sets) attributes.
- We illustrate the use of methodology on a sample survey with 23 direct and 9 indirect questions applied to 656 respondents.
- We derive actionable business rules using a decision-tree-based classification model. These results can be conveniently integrated into the decision support systems of financial institutions.

In this section of the chapter, the study was introduced and motivated. In Section 2, an overview of the basic concepts in related studies is presented through a concise literature review. In Section 3, the proposed five-step methodology is presented and the methodology steps are illustrated through a sample survey study. In Section 4, the study is concluded with a thorough discussion of future research directions.

2 Literature

A number of researchers have evaluated the use of risk surveys by financial institutions to score the risk preferences of individuals and to classify them into categories. [26]'s

evaluation of 14 risk surveys (questionnaires) used in France finds that only one third of the surveys try to quantify risk aversion, and those who quantify risk aversion fail to use sufficiently sophisticated econometric methods. Less than half of the institutions have developed scoring rules for the purpose of classification, and for most cases, classification is conducted based on subjective judgment rather than proper analytical methods. In addition, computed classes are only weakly correlated between different surveys. Our study addresses some of these issues.

Researchers have long discussed whether indirect attributes can be used effectively in classifying individuals into risk preference categories. See, for example, [13]. In particular, being male, being single, being a professional employee, younger age, higher income, higher education, higher knowledge in financial matters and having positive economic expectations are shown to be positively related to higher risk tolerance. However, blindly adopting such heuristics in classifying customers has its drawbacks. There is no consensus among researchers about the validity of these heuristics, which indicate the need for additional research (see [14] and the references therein). For example, using a survey with 20 questions, [13] finds older individuals to be more risk tolerant than younger ones, and married individuals to be more risk tolerant than single ones, which contradicts the common expectations. In a similar study, [14] uses the 1992 Survey of Consumer Finances (SCF) dataset, which contains the answers of 2626 respondents. Seven of the eight indirect attributes are found to be effective in classifying respondents into three risk tolerance categories. *The level of attained education* and *gender* are found to be the most effective attributes; whereas the effect of *age* attribute is found to be insignificant. Other related studies include [31] and [16].

A different but related problem is the *credit scoring* problem. Credit scoring can be defined as the application of quantitative methods to "predict the probability that a loan applicant or existing borrower will default or become delinquent" [22]. Credit scoring models are popular in finance, due to increasing competition in the industry and the high cost of bad debt. [11] presents a review of credit scoring models based on statistical techniques and learning techniques, and their applications. [33] provides a review of credit scoring and behavior scoring models, where the latter type of models use data on the repayment and ordering history of a given customer. Numerous novel credit scoring models have been published after the reviews of [11, 33], and are based on a variety of techniques; including neural networks [34], self-organizing maps [18], feature selection, case based reasoning, support vector machines (SVM) [19], discriminant analysis, multivariate adaptive regression splines (MARS), clustering, and combinations of these techniques [6]. [30] develops a credit scoring framework and an expert system based on neuro-fuzzy logic to assess creditworthiness of an entrepreneur.

Different from our study, the mentioned studies do not focus on the individual's attitude towards risk, namely, his/her risk preference. Risk preference and being risky from a lender's perspective are different issues. For example, an individual who is very much risk-seeking may or may not have a high credit score (low credit risk). Also, these studies do not provide an algorithm for determining scores in the absence of a learning set.

Another research stream consists of the literature on customer segmentation as a part of Customer Relationship Management (CRM). [24] presents a summary of the

research on supervised classification for CRM. [20] employs decision tree models for not only generating business rules regarding behavior patterns of customers, but also for dynamically tracking the changes in these rules.

We develop a numerical score for representing risk preferences with regards to financial decision making, using data from a field survey. However, risk preferences can also be estimated through controlled field experiments [17]. These experiments often identify deviations in human behavior from theoretical predictions, which is studied in the *behavioral finance* literature [4].

3 Methodology and Results

Our methodology is outlined below. In the following subsections, each step of the methodology is presented alongside the results we obtain based on our sample survey data.

1. Survey design
2. Survey conduct
3. Risk scoring
4. Classification
5. Insight generation

3.1 Survey Design

We investigated the risk scoring surveys of a number of financial institutions available on the Web, and developed our survey by choosing 23 direct (risk-related) and 9 indirect (demographic) questions among the popular ones. Appropriate selection of the direct attributes for a survey directly affects the risk scores and the subsequent data mining study, and hence is very important.

The questions in the survey were designed such that the *choices* given to respondents are sorted according to (hypothesized) risk preferences. For example, in the survey questions with three choices, selecting choice (a) is assumed to reflect risk-averse behavior, whereas selecting choice (c) is assumed to reflect risk-seeking behavior.

Examples of the questions on direct attributes include the number of times a person plays in the stock market, the investment types that a person would feel more comfortable with, and the most important investment goal of that person. A number of sample direct questions is provided in Appendix A. The complete survey (English version) is provided in Appendix A of the supplementary document for this chapter [10].

We used the following nine indirect attributes in our study:

- *Gender*: male or female
- *IsStudent*: whether the person an employee or a student
- *StudentLevel*: undergrad, masters
- *IncomeType*: fixed salary, incentive based, or both
- *SoccerTeam*: the soccer team that the person supports
- *HighschoolType*: public, private, public science, private science, other

– *EnglishLevel*: the level of English language skill
– *GermanLevel*: the level of German language skill
– *FrenchLevel*: the level of French language skill

The other indirect questions in the survey, such as the department that a student studies in, were not included in the scoring and prediction phases of the study because they were open-ended.

3.2 Survey Conduct

The survey was conducted in Turkish language on 656 respondents, with balanced distribution of working people (346) vs. students (250 undergraduates and 60 graduates), and gender (283 females vs. 373 males), from a multitude of universities and work environments. Among the working participants, 71 work only for commission, 204 work for fixed income and 71 work for both commission and fixed income. The distribution of values for the attributes are given in Appendix B of the supplementary document [10].

One challenge faced while conducting the survey was the communication of finance and insurance concepts, and the choices available to respondents. This is important for ensuring valid answers to the survey questions and hence improving the reliability of the sample study. To this end, all surveys were conducted through one-to-one interaction with individuals by our research team. One drawback of this approach is that communication may influence respondents' risk preferences. For example, [27] observes that farmers in Netherlands exhibit more risk-seeking behavior when they understand and trust the insurance tools through one-to-one interaction. The results of [27] confirm earlier findings in India, Africa, and South America. We do not analyze the effects of such a bias.

Once the survey was conducted, the data was assembled in a spreadsheet software and cleaned following the guidelines in the taxonomy of dirty data in [21]. Also at the data cleaning stage, data was anonymized, so that it can be shared with colleagues and students in future projects.

3.3 Risk Scoring

The survey data is fed into the *risk scoring* algorithm in the form of an $I \times J$ sized matrix, representing I respondents and J attributes. This algorithm determines which direct attributes are to be used in scoring, the weights for each attribute, and based on these, the risk scores for each respondent. The mathematical notation and the pseudocode of the scoring algorithm are given in Appendix B.

The initialization step in the algorithm linearly transforms ordinal choice data into nominal values between 0 and 3. For example, if a question has five choices (a, b, c, d, e), the corresponding numerical values would be (0.00, 0.75, 1.50, 2.25, 3.00). This linear transformation is used for simplicity; however, there is no guarantee it is the most accurate representation.

Following the initialization phase, quantitative attribute values are fed into a regression-based iterative algorithm. The algorithm operates as a multi-pass self-organizing heuristic, which aims at obtaining converged risk scores. The stopping criterion is satisfied

when the average absolute percentage difference in risk scores is less than the threshold provided by the analyst. At each iteration of the algorithm, the value vector for each of the selected attributes is entered into a linear regression model as factor, where the response is the incumbent risk score vector. Weights for the attributes are updated at the beginning of each iteration, such that the sum of the weights is equal to the number of included attributes. The algorithm allows for change in the direction of signs when the choices for an attribute should take decreasing -rather than increasing- values from choice (a) to the final choice. Hence, the algorithm not only eliminates irrelevant attributes, but also suggests the direction of risk preferences for the choices of a given attribute. The algorithm is an unsupervised algorithm, as it does not require any class labels or scores from the user. It is also a self-organizing algorithm [2], as it automatically converges to a solution at the desired error threshold.

After the risk scores are calculated for all respondents, a certain top percentage of them are labeled as *risk-seeking* and the rest as *risk-averse*. This is used in the subsequent *classification* step of the methodology.

The algorithm was coded in Matlab computational environment [25]. The mapping of the ordinal values $\mathbf{O} = [o_{ij}]_{656\times 23}$ to nominal values in the initialization step was performed in the spreadsheet software, and the Matlab code was run with the obtained matrix of nominal values $\mathbf{A} = [a_{ij}]_{656\times 23}$. The parameters for the algorithm were selected as $E = 0.1$ and $\alpha = 0.05$. Running time for the algorithm was negligibly small (less than one second) for this sample.

Results on scoring algorithm:

The average absolute percentage change $\overline{e}_k$ in risk scores is shown in Fig. 1. We observe $\overline{e}_k$ to halve in only two iterations, and to get very close to zero after the first 10 iterations. The algorithm converges to the given threshold E rapidly, in only 19 iterations.

Fig. 2 shows the weights obtained for each of the 23 direct attributes. Five of the 23 direct attributes (Q20, Q21, Q22, Q38, Q40) are assigned a weight of 0 by the algorithm. That is, the algorithm removes these five questions from the risk score computations, because they fail to impact the scores in a statistically significant way, given the presence of the other 18 attributes. The positive weights are observed in the range (0.2792,

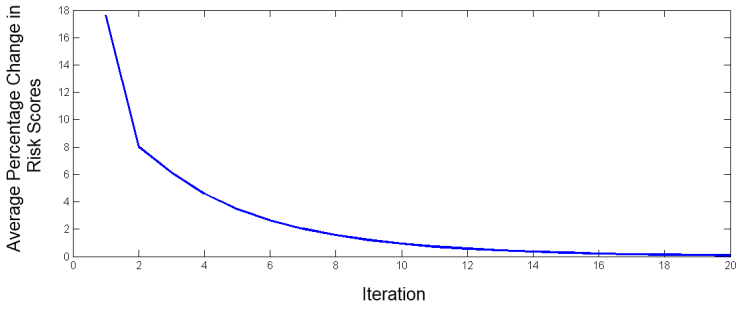

Fig. 1. The convergence of the algorithm, based on the average percentage change in risk scores

1.6320). The hypothesized directions of choice ranks are found to be correct for all the attributes ($\Gamma_j = 1, \forall j \in \mathcal{J}$).

Fig. 3 illustrates the histogram of the risk scores we calculate, labeling 20% of the respondents as risk-seeking and the rest as risk-averse. While the risk scores seem to exhibit normal distribution, Shapiro-Wilk test for normality [29], carried out in R statistical package [32], resulted in $p = 3.2E-7 \ll 0.05$, very strongly suggesting a non-normal fit.

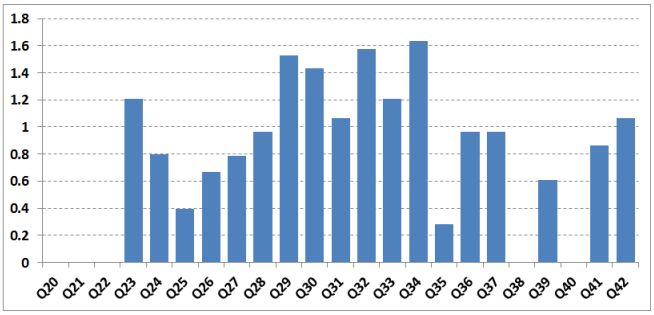

Fig. 2. Calculated weights for the direct attributes in the case study

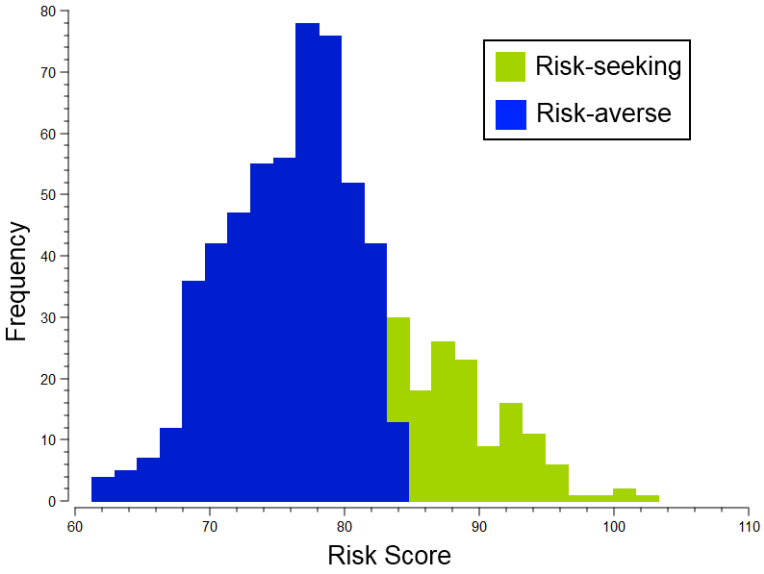

Fig. 3. Risk score histogram and the definition of (risk) class labels

3.4 Classification

The next step in the methodology investigates whether risk preferences can be predicted through only direct, only indirect or both sets of attributes. To this end, we use five *classification algorithms* from the field of machine learning for predicting whether a person is risk-seeking or risk-averse, as labeled by the scoring algorithm of step 3. These algorithms, also referred to as *learners*, are Naive Bayes, k-Nearest Neighbor (kNN), C4.5, Support Vector Machines (SVM), and Decision Trees (DT)[4] [8].

In classification models, a *learning dataset* is used by the learner for supervised learning to later on predict the class label for new respondents. The predictors can have nominal or categorical values, whereas the predicted class attribute should have categorical (class label) values. The success of a learner is measured primarily through *classification accuracy* on a provided *test dataset*, besides a number of other metrics. Classification accuracy is defined as the percentage of correct predictions made by the classification algorithm on the test dataset.

Fig. 4 illustrates the generic classification model we construct for risk preference prediction, as well as the widgets for decision tree analysis in the Orange data mining software [35]. In the classification model, some of the attributes in the full dataset are selected as the predictors and the risk-preference attribute (taking class label values of *risk-seeking* or *risk-averse*) is selected as the class attribute.

Classification accuracy is computed through 70% sampling with ten repeats. In other words, for each learner, ten experiments are carried out, with a random 70% of

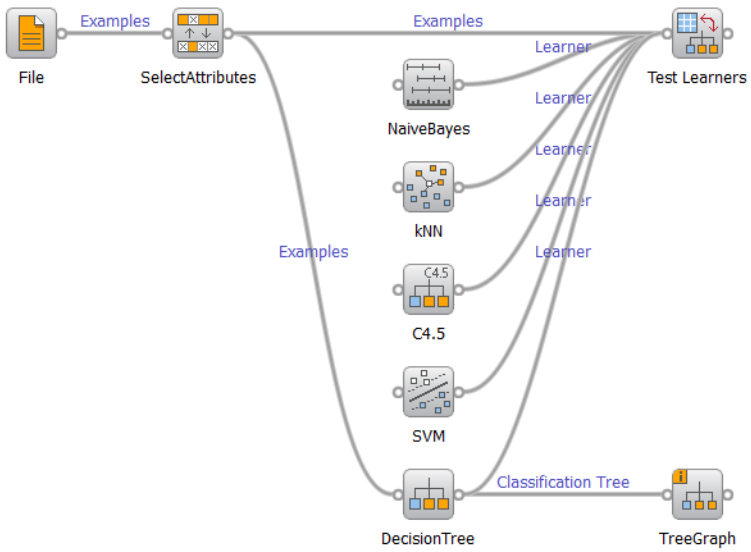

Fig. 4. Classification model for predicting risk preference behavior, together with the decision tree analysis widgets.

[4] also referred to as *classification trees*

Table 1. Classification accuracies of the models for predicting risk preferences

	Learner	Model 1a	Model 2a	Model 3a	Model 1b	Model 2b	Model 3b
1	Naive Bayes	0.9635	0.7888	0.9650	0.9467	0.8954	0.9452
2	kNN	0.9279	0.7675	0.9091	0.9391	0.8756	0.9452
3	C4.5	0.9279	0.8020	0.9269	0.9198	0.8985	0.9173
4	SVM	0.9528	0.8020	0.9452	0.9650	0.8985	0.9584
5	DT	0.9142	0.7741	0.9030	0.9239	0.8919	0.9137

the sample being used as the training dataset each time, and the remaining used as the test dataset.

Results on classification models:

Table 1 presents the classification accuracy results of the six models. We observe that in Model 1a, Naive Bayes learner successfully classifies (on the average) 96.35% of the respondents in the test dataset. This is not surprising, since the direct attributes that Model 1a uses were used in the computation of risk scores in the first place, which are eventually transformed into the risk preference class labels. Therefore, high classification accuracy for Model 1a is expected. What is surprising is the relatively high (around 80%) classification accuracy that the learners in Model 2b achieve. This finding suggests that indirect attributes can be almost as successful as direct attributes in predicting risk preference.

Another surprising outcome is the poor performance of Models 3a and 3b, which use both direct and indirect attributes. Model 3a is outperformed by Model 1a with all but one learner. The comparison between Models 3b and 1b is also similar. This observation suggests that if one is already using the direct attributes, adding indirect attributes can deteriorate the classification performance of learners.

While not yielding the highest classification accuracy in any of the models, decision tree (DT) may be preferred over other (black-box) learners due to its strong explanatory capacity, in the form of explicit rules it generates. We discuss one decision tree application in the following step.

3.5 Insight Generation

In this step of the methodology, we aim to determine whether the answers to direct or indirect questions convey information about the risk preferences of respondents. To this end, a decision tree is constructed in the Orange model.

Decision trees summarize rule-based information regarding classes using trees. As opposed to the black-box operation of machine learning algorithms, decision trees return explicit rules, in the form *"IF Antecedent THEN Consequent"*, that can easily be understood and adopted for real world applications. For example, in the context of risk, [23] gives an example rule which states that credit card holders who withdrew money

at casinos had higher rates of delinquency and bankruptcy. Such rules can also encapsulate the domain knowledge in expert systems development, in the form of *rule bases* [12]. Wagner et al. [36] state that knowledge acquisition is the greatest bottleneck in the expert system development process, due to unavailability of experts and knowledge engineers and difficulties with the rule extraction process. Our methodology offers a recipe for this important bottleneck of expert systems development.

In decision trees, branching is carried out at each node according to a *split criterion* and a tree with a desired depth is constructed. At each deeper level, the split that yields the most increase in the split criterion is selected. [7] gives a concise review of algorithms for decision tree analysis, explaining the characteristics of each algorithm. In our decision tree analysis, we use the ID3 algorithm [28] in Orange software [35] that creates branches based on the *information gain* criterion. Each level in the decision tree is based on the value of a particular variable. For instance, in Fig. 5, the *root node* of the decision tree contains 656 respondents and the branching is based on question 34 ($Q34$). In each node of this decision tree, the dark slice of the pie chart shows the proportion of risk-seeking participants, and the remaining portion of the pie shows risk-averse respondents in that sub-sample. In decision trees, we are especially interested in identifying the nodes that differ significantly from the root node with respect to the shares of the slices, and the splits that result in significant changes in the slices of the pie chart compared to the *parent node* (the node above the split).

Fig. 5 shows the decision tree for Model 1a, where only the direct attributes are used. We observe a significant branching based on the answer given to $Q34$ (question 34). When *Q34* takes the value *a*, the percentage of risk-seeking respondents drops significantly from 20.00% (Definition "a") to just 1.53% (2 out of 130 respondents).

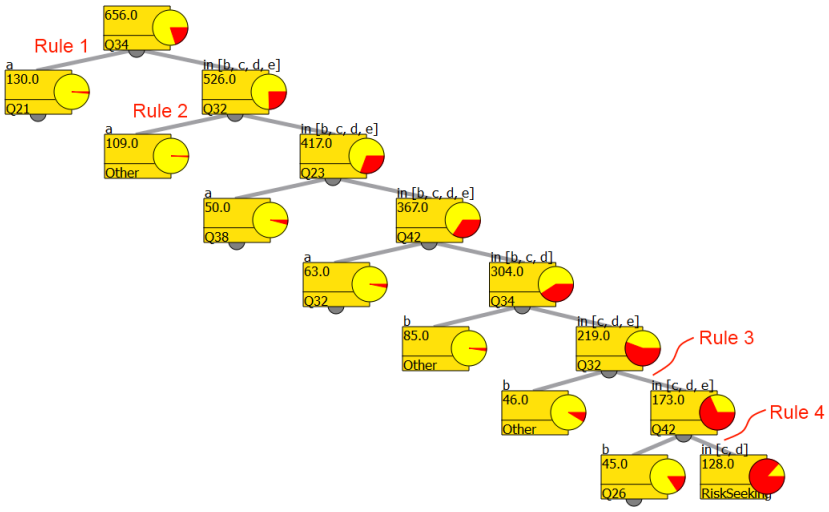

Fig. 5. Decision tree for Model 1a, where only direct attributes are used and the respondents with the top 20% highest scores are labeled as risk-seeking

Similarly, even if $Q34$ takes a value in $\{b,c,d,e\}$, if $Q32$ has the value of a, then again, the chances of that person being risk-seeking in this sample is much lower (actually 1, out of 109 respondents) than the root node (that represents the complete sample). Rule 1 and Rule 2, labeled on Fig. 5, reflect the aforementioned findings as below:

Rule 1: "**IF** $Q34 = a$ **THEN** $Proportion(RiskSeeking) = 1.53\%$."

Rule 2: "**IF** $Q34 \in \{b,c,d,e\} \wedge Q32 = a$ **THEN** $Proportion(RiskSeeking) = 0.92\%$."

As $Q34, Q32$, and $Q23$ (questions with five choices) take values in $\{b,c,d,e\}$ (any value but a), and $Q42$ (a question with four choices) takes a value in $\{b,c,d\}$, the proportion of the risk-seeking respondents continues to increase compared to the root node. The next three splits are again related with these questions, and hence these four questions are the most important risk-related questions when deriving rules for Model 1a. $Q34$ and $Q32$ also had the largest weights in the scoring algorithm (as seen in Figure 2), but $Q23$ and $Q42$ did not have the next two largest weights. This tells us that the weights obtained by the scoring algorithm are related, but not perfectly aligned with the results of the decision tree analysis. These questions ask about the volatility level that the person would be willing to accept ($Q34$), top investment priority ($Q32$), a self-assessment of risk preference compared to others ($Q23$), and the most preferred investment strategy ($Q42$).

The rules that corresponds to the splits marked with Rule 3 and Rule 4 in Fig. 5 are as follows:

Rule 3: "**IF** $Q34 \in \{c,d,e\} \wedge Q32 \in \{c,d,e\} \wedge Q23 \in \{b,c,d,e\} \wedge Q42 \in \{b,c,d\}$ **THEN** $Proportion(RiskSeeking) = 68.21\%$."

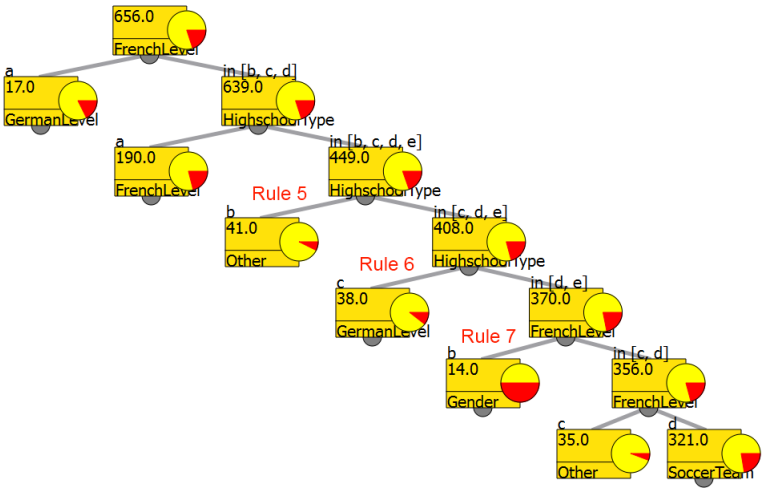

Fig. 6. Decision tree for Model 2a, where only indirect attributes are used and the respondents with the top 20% highest scores are labeled as risk-seeking

Rule 4: "**IF** $Q34 \in \{c,d,e\} \wedge Q32 \in \{c,d,e\} \wedge Q23 \in \{b,c,d,e\} \wedge Q42 \in \{c,d\}$ **THEN** $Proportion(RiskSeeking) = 86.72\%$."

The only difference between Rules 3 and 4 is that in Rule 4, Q42 takes values of c or d, rather than a value in $\{b,c,d\}$.

Next, we discuss the decision tree for Model 2a, where only the indirect attributes are used. As the decision tree in Fig. 6 suggests, $FrenchLevel$ and $HighschoolType$ are the attributes that result in the most fundamental splits. A significant change takes place in the pie structure when $HighschoolType = b$ (public science high school[5]), given that $FrenchLevel \in \{b,c,d\}$. Specifically, in the mentioned split, the pie slice that corresponds to risk-seeking respondents becomes much smaller (3 respondents out of 41) compared to its parent node. This is Rule 5, labeled in Fig. 6 and stated below.

Rule 5: "**IF** $FrenchLevel \in \{b,c,d\} \wedge HighschoolType = b$ **THEN** $Proportion(RiskSeeking) = 7.31\%$."

A similar split takes place when $HighschoolType = c$ (private science high school), again resulting in a low proportion (4 out of 38) of risk-seeking respondents:

Rule 6: "**IF** $FrenchLevel \in \{b,c,d\} \wedge HighschoolType = c$ **THEN** $Proportion(RiskSeeking) = 10.53\%$."

It is striking that respondents who graduated from public and private science high schools are much more risk-averse, compared to other sub-groups. This has important implications for the business world: Our finding suggests that it is unlikely for respondents with a strong science background in high-school to establish risky businesses, such as high-technology startups. However, such startups are highly critical in the development of an economy, and are dependent on the know-how of technically competent people, such as professionals with a strong science background beginning in high school. Therefore, there should be mechanisms to encourage risk-taking behavior among science high school alumni, and to establish connections between graduates of science high schools and those with an entrepreneurial mindset.

In Fig. 6, another major split takes place in the split labeled as Rule 7. Here, the question that creates the split is $FrenchLevel = b$ (French level is intermediate), given that $HighschoolType \in \{d,e\}$ (private high school, and state high schools with foreign language). $FrenchLevel = b$ results in a very high proportion (7 out of 14) of risk-seeking respondents. Yet, the number of respondents in the mentioned sub-sample is very small, only 14, and this rule has to be handled with caution.

Rule 7: "**IF** $HighschoolType \in \{d,e\} \wedge FrenchLevel = b$ **THEN** $Proportion(RiskSeeking) = 50.00\%$."

Upon further querying, we find that only 3 out of 17 respondents (17.65%) that have $FrenchLevel = a$ are risk-seeking, whereas 8 out of 18 respondents (44.44%) that have $FrenchLevel = b$ are risk-seeking. Risk-seeking behavior is minimal (9.33%) among the 75 respondents that have $FrenchLevel = c$. What could be the explanation for such

[5] *Science high school*: specially designated high schools that heavily implement a math- and science-oriented curriculum

a pattern? One possible explanation might be the following: In Turkey, individuals that have $FrenchLevel = a$ typically come from wealthy families and learn French in expensive private high schools. Individuals that have $FrenchLevel = b$ typically have strived to learn French by themselves, without going to such schools. They have aspirations to rise socio-economically, and are ready to take the risks needed to achieve their aspirations. Definitely, the true explanation for this pattern is a research question for the field of sociology.

4 Conclusions

In this study, we develop a scoring algorithm, implement it with real world survey data, and obtain significant insights through mining risk scores for the sample. In particular, we find that demographic attributes of individuals can be used to predict their risk preference categories. This result has important practical implications: Without asking any risk-related questions, but by only obtaining demographic information, one can estimate with reasonable accuracy whether a particular respondent is risk-seeking or risk-averse. The data for those indirect attributes is often routinely collected on the Internet when registering for web sites. This would eliminate the need to collect extensive finance-related information or sensitive personal information [37] from customers. Another advantage is that, respondents would typically not distort their answers to indirect questions, whereas they could do so with direct ones. Hence, our methodology can feasibly be implemented in practice, and has the potential to bring significant predictive power to the institution at minimal effort.

Classification of customers into risk preference categories is an important problem for financial institutions. As argued in [14], incorrectly classifying a risk-averse customer as risk-seeking may later cause the customer to sell investments at a loss; whereas the opposite mistake may cause the customer to miss his investment objectives. Using our methodology, the institution can make a pre-classification of customers into risk-averse and risk-seeking categories. If necessary, these customers can then be given surveys with more tailored direct (risk-related) questions. The computational nature of our methodology makes it easy to be integrated into existing CRM systems in terms of data use and result feed.

The methodology and the scoring algorithm proposed in this work are actually *platforms* on which better methodologies and algorithms can be designed. There exists a rich possibility of future research on this area, mostly regarding the algorithm:

- The algorithm assumes that the risk score of each respondent can be computed with the same set of attribute weights. However, different weights may apply to different subgroups within the population. This can be analyzed by incorporating cluster analysis [38, 39] into the current study.
- The numeric values assigned to the ordinal values of the attributes were assumed to be linear and equally spaced; whereas the real relation may be highly nonlinear. Linearizable functions [9] or higher order polynomials can be assumed for attributes as a whole, or each attribute may be modeled flexibly to follow any of these functional forms. As an even more general model, weights can be computed not only for attributes, but for each choice of each attribute.

– In scoring, statistical techniques for feature selection and dimensionality reduction that exist in literature [15] may be adopted to obtain approximately the same results with fewer direct questions. This problem can be solved together with the outlier detection problem, as in [1], where the authors present a hybrid approach combining case-based reasoning (CBR) with genetic algorithms (GAs) to optimize attribute weights and select relevant respondents simultaneously.
– The scoring algorithm can be developed such that consistent results are obtained for different samples. For example, in the ideal case, a respondent who answered the same question in a particular sample should have same score if he was a part of another sample. In our presented algorithm, each respondent's risk score is dependent on the answers of the whole sample. This will not pose a problem when the methodology is applied to large data sets, such as all customers of a financial institution.
– The proposed methodology eliminates irrelevant *direct* attributes in computing the risk scores, but it does not eliminate *indirect* attributes that are irrelevant or do not provide significant information. All the potential indirect attributes are considered in the classification models. Dimensionality reduction techniques can be used in this step of the methodology. This would allow asking as few indirect questions as possible, but still being able to predict risk preference with a high accuracy.

Acknowledgement

The authors thank Sabancı University (SU) alumni Levent Bora, Kıvanc Kılınç, Onur Özcan, Feyyaz Etiz for their work on earlier phases of the study, and students Serpil Çetin and Nazlı Ceylan Ersöz for collecting the data for the case study. The authors also thank SU students Gizem Gürdeniz, Havva Gözde Ekşioğlu and Dicle Ceylan for their assistance. This chapter is dedicated to the memory of Mr. Turgut Uzer, a leading industrial engineer in Turkey, who passed away in February 2011. Mr. Turgut Uzer inspired the authors greatly with his vision, unmatched know-how, and dedication to the advancement of decision sciences.

Appendix A: Selected Survey Questions

Following are selected direct (risk-related) questions from the survey of the case study, which constitute the corresponding direct (risk-related) attributes.

Q34. Over the long term, typically, investments which are more volatile (i.e., that tend to fluctuate more in value) have greater potential for return (Stocks, for example, have high volatility; whereas government bonds have low volatility). Given this trade-off, what would be the level of volatility you would prefer for your investment?

a Less than 3%
b 3% to 5%
c 5% to 7%
d 7% to 13%
e More than 13%

Q32. What is your most important investment priority?

a I aim to protect my capital; I cannot stand losing money.
b I am OK with small growth; I cannot take much risk.
c I aim for an investment that delivers the market return rate.
d I want higher than market return; I am OK with volatility.
e Return is the most important for me. I am ready to take high risk for high return.

Q23. Compared to others, how do you rate your willingness to take risk?

a Very low
b Low
c Average
d High
e Very high

Q42. What is your most preferred investment strategy?

a I want my investments to be secure. I also need my investments to provide me with modest income now, or to fund a large expense within the next few years.
b I want my investments to grow and I am less concerned about income. I am comfortable with moderate market fluctuations.
c I am more interested in having my investments grow over the long-term. I am comfortable with short-term return volatility.
d I want long-term aggressive growth and I am willing to accept significant short-term market fluctuations.

Appendix B: ScoringAlgorithm

Following is the mathematical presentation of the developed scoring algorithm:

Sets

$\mathcal{I}$: set of respondents (observations, rows) in the sample; $i = 1, \cdots, I$

$\mathcal{J}$: set of attributes (questions, columns); $j = 1, \cdots, J$

$\mathcal{V}$: set of ordinal values for each attribute; $v = 1, \cdots, V$. For the presented case study, $\mathcal{V} = (a, b, c, d, e)$, where $a \leq b \leq c \leq d \leq e$

Inputs

$\mathbf{O} = [o_{ij}]_{I \times J}$: matrix of ordinal values of all attributes for all respondents

m_j: number of possible ordinal values for attribute j; $m_j \leq 5$ in this study

Internal Variables

$\mathbf{A} = [a_{ij}]_{I \times J}$: matrix of numerical (nominal) values of all attributes for all respondents

y_i: temporary adjusted risk score for respondent i, to be used in regression

Parameters

E: threshold on absolute percentage error (falling below this value will terminate the algorithm)

α: threshold for type-1 error (probability of rejecting a hypothesis when the hypothesis is in fact true)

M: a very large number

$\mathbf{B}$: transformation matrix for converting the ordinal input value matrix $\mathbf{O}$ into the numerical (nominal) value matrix $\mathbf{A}$

Outputs

z_j: whether attribute j is to be included in computing the risk score; $z_j \in \{0,1\}$

w_j: weight for attribute j; $w_j \geq 0$

β_{0j}: intercept value for attribute j

β_{1j}: slope value for attribute j

Γ_j: sign multiplier for attribute j; $\Gamma_j \in \{-1,1\}$

x_i: risk score for respondent i

Functions

$f(v,n) : (\mathcal{V}, \{2,\cdots,V\}) \rightarrow [0,3]$: mapping function for an attribute with n possible values, that transforms the ordinal value v collected for that attribute to a nominal value $b_{v,n-1}$.

$$f(v,n) = b_{v,n-1}$$

where, for $V = 5$,

$$\mathbf{B} = [b_{vn}]_{V\times(V-1)} = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 \\ 3.00 & 1.50 & 1.00 & 0.75 \\ \cdot & 3.00 & 2.00 & 1.50 \\ \cdot & \cdot & 3.00 & 2.25 \\ \cdot & \cdot & \cdot & 3.00 \end{bmatrix}$$

regression$(\mathbf{y}, \mathbf{a}')$

solve regression model $\mathbf{y} = \beta_0 + \beta_1 \mathbf{a}' + \varepsilon$ for vectors $\mathbf{y}$ and $\mathbf{a}'$
return $(p,\ \beta_0, \beta_1)$, where p is the p-value for the regression model

preprocess()

// transform ordinal attribute values to nominal values
$a_{ij} = f(o_{ij}, m_j); \forall (i,j) \in \mathcal{I} \times \mathcal{J}$

Iteration-Related Notation

k: iteration count

N: number of attributes included in risk score computations at a given iteration

W: sum of weights for attributes

ε_k: absolute error at a given iteration k

e_k: absolute percentage error at a given iteration k

$\overline{e}_k$: average absolute percentage error at a given iteration k

ScoringAlgorithm $(\mathbf{O},\ m_j)$

BEGIN

// perform pre-processing to transform ordinal data to nominal data

preprocess()

// initialization:

// initially, all attributes are included in scoring,

// with unit weight of 1 and sign multiplier of 1.

// all of the regression intercepts are 0.

$z_j = 1,\ w_j = 1,\ \Gamma_j = 1,\ \beta_{0j} = 0; \forall j \in \mathcal{J}$

$N = \sum_j z_j$

// begin with iteration count of 1

$k = 1$

Begin_Iteration

// standardize the weights, so that their sum W will equal to N

$W = \sum_j w_j z_j$

$w_j \leftarrow (N w_j)/W;;\ \forall j \in \mathcal{J}$

// compute the average of the intercepts

$\overline{\beta}_{0\cdot} = (\sum_j \beta_{0j} z_j)/N$

// compute/update the risk scores at iteration k,

// which is composed of the average intercept value

// and the sum of weighted values for attributes

$x_{ik} = \overline{\beta}_{0\cdot} + \sum_j \Gamma_j\ w_j a_{ij}; \forall i \in \mathcal{I}$

// compute total absolute error

$\varepsilon_k = \sum_i \left| x_{ik} - x_{i,k-1} \right|$

// correction for the initial error values

if $k = 1$ **then**

$\varepsilon_0 = \varepsilon_1$

// termination condition

if $\varepsilon_k = 0$ **then**

go to $Iterations\_Completed$

// compute absolute percentage error,

// and then its average over the last two iterations

$\bar{x}_{\cdot k} = \sum_i x_{ik}/I$

$e_k = 100\varepsilon_k/\bar{x}_{\cdot k}$

$\bar{e}_k = (e_k + e_{k-1})/2$

// if the stopping criterion is satisfied, terminate the algorithm

if $\bar{e}_k < E$ **then**

go to $Iterations\_Completed$

// otherwise, continue with the regression modeling for each attribute j,

// and then go to next iteration

$\forall j \in \mathcal{J}$

// if the attribute is included in the risk score calculation

if $z_j = 1$ **then**

// first remove the attribute value from the incumbent score

// to eliminate its effect

$y_i = x_{ik} - a_{ij};\ \forall i \in \mathcal{I}$

// then define the vectors for the regression model of that attribute

$\mathbf{y} = (y_i)\,; \mathbf{a}' = (\Gamma_j\ a_{\cdot j})$

$(p,\ \beta_0, \beta_1) = regression(\mathbf{y}, \mathbf{a}')$

// if the regression yields a high p value

// that is greater than the type-1 error,

// this means that attribute j does not contribute significantly

// to the risk scores

if $p > \alpha$ **then**

// and the attribute should not be included in risk calculations

$z_j = 0$

else

// else it will be included (will just keep its default value)

$z_j = 1$

// and weight for the attribute will be the slope value

// obtained from the regression

$w_j = \beta_1$

// the sign of the slope is important;

// if it is negative, this should be noted

if $\beta_1 < 0$ **then**

// record the sign change in the sign multiplier

$\Gamma_j = -1$

else

$\Gamma_j = 1$

// advance the iteration count and begin the next iteration

$k++$

go to *Begin_Iteration*

Iterations_Completed

$x_i = x_{ik}$

return $x_i,\ z_j, w_j, \Gamma_j,\ \beta_{0j}$

END

References

1. H. Ahn, K. Kim, and I. Han. Hybrid genetic algorithms and case-based reasoning systems for customer classification. *Expert Systems*, 23(3):127–144, 2006.
2. W.R. Ashby. Principles of the self-organizing system. *Principles of Self-organization*, pages 255–278, 1962.
3. T. Aven and O. Renn. *Risk Management and Governance: Concepts, Guidelines and Applications*. Springer Verlag, 2010.
4. N. Barberis and R.H. Thaler. *A survey of behavioral finance*, in *Handbook of the economics of finance*, Volume 1, Part 1, George M. Constantinides, Milton Harris, René M. Stulz (Eds.). pages 1053–1128. Elsevier, 2003.
5. L. Cao. Behavior informatics and analytics: Let behavior talk. In *ICDMW '08. IEEE International Conference on Data Mining Workshops, 2008.*, pages 87–96, 2008.
6. F.L. Chen and F.C. Li. Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, 37(7):4902–4909, 2010.
7. C.F. Chien and L.F. Chen. 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, 2008.
8. B. Clarke, E. Fokoué, and H.H. Zhang. *Principles and theory for data mining and machine learning*. Springer Verlag, 2009.

9. C. Daniel and F.S. Wood. Fitting functions to data. New York: Wiley, 1980.
10. G. Ertek, M. Kaya, C. Kefeli, C. Onur, and K. Uzer. Supplementary document for "Scoring and Predicting Risk Preferences", Available online under `http://people.sabanciuniv.edu/ertekg/papers/supp/03.pdf`. 2011.
11. J. Galindo and P. Tamayo. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1):107–143, 2000.
12. J.C. Giarratano and G. Riley. *Expert systems: principles and programming*. Brooks/Cole Publishing Co., 1989.
13. J. E. Grable. Financial risk tolerance and additional factors that affect risk taking in everyday. *Journal of Business and Psychology*, 14(4):625–630, 2000.
14. J.E. Grable and R.H. Lytton. Investor risk tolerance: Testing the efficacy of demographics as differentiating and classifying factors. *Financial Counseling and Planning*, 9(1):61–74, 1998.
15. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
16. T.A. Hallahan, R.W. Faff, and M.D. Mckenzie. An empirical investigation of personal financial risk tolerance. *Financial Services Review*, 13(1):57–78, 2004.
17. G.W. Harrison, M.I. Lau, and E.E. Rutstrom. Estimating risk attitudes in Denmark: A field experiment. *Scandinavian Journal of Economics*, 109(2):341–368, 2007.
18. N.C. Hsieh. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, 27(4):623–633, 2004.
19. C.L. Huang, M.C. Chen, and C.J. Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847–856, 2007.
20. J.K. Kim, H.S. Song, T.S. Kim, and H.K. Kim. Detecting the change of customer behavior based on decision tree analysis. *Expert Systems*, 22(4):193–205, 2005.
21. W. Kim, B.J. Choi, E.K. Hong, S.K. Kim, and D. Lee. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1):81–99, 2003.
22. H.C. Koh, CT Wei, and PG Chwee. A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information*, 1(1):96–118, 2006.
23. L. Kuykendall. September 1999. The Data-Mining Toolbox. *Credit Card Management*, 12(7).
24. S. Lessmann and S. Voß. Supervised classification for decision support in customer relationship management, in *Intelligent Decision Support*, by Andreas Bortfeldt (Ed.), page 231, 2008.
25. MathWorks. Matlab, `http://www.mathworks.com`. 2011.
26. A. Palma and N. Picard. Evaluation of MiFID questionnaries in France. Technical report, AMF, 2010.
27. A. Patt, N. Peterson, M. Carter, M. Velez, U. Hess, and P. Suarez. Making index insurance attractive to farmers. *Mitigation and Adaptation Strategies for Global Change*, 14(8):737–753, 2009.
28. J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
29. S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.
30. D.K. Sreekantha and R.V. Kulkarni. Expert system design for credit risk evaluation using neuro-fuzzy logic. *Expert Systems*. doi: 10.1111/j.1468-0394.2010.00562.x.
31. J. Sung and S. Hanna. Factors related to risk tolerance. *Financial Counseling and Planning*, 7, 1996.
32. The R Foundation for Statistical Computing. R Project, `http://www.r-project.org`. 2011.

33. L.C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172, 2000.
34. C.F. Tsai and J.W. Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4):2639–2649, 2008.
35. University of Ljubljana, Bioinformatics Laboratory. Orange, `http://orange.biolab.si/`. 2011.
36. W.P. Wagner, M.K. Najdawi, and Q.B. Chung. Selection of knowledge acquisition techniques based upon the problem domain characteristics of production and operations management expert systems. *Expert Systems*, 18(2):76–87, 2001.
37. X.T. Wang, D.J. Kruger, and A. Wilke. Life history variables and risk-taking propensity. *Evolution and Human Behavior*, 30(2):77–84, 2009.
38. R. Xu, and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.
39. D. Zakrzewska and J. Murlewski. Clustering algorithms for bank customer segmentation. 2005.