Identifying Exceptional Descriptions of People using Topic Modeling and Subgroup Discovery

Andrew T. Hendrickson, Jason Wang, and Martin Atzmueller

Tilburg University, 5037AB, the Netherlands
{a.hendrickson, y.w.wang, m.atzmuller}@uvt.nl

Abstract. Descriptions of images form the backbone for many intelligent systems that identify images, generate descriptions of novel images, or cluster and categorize images based on tags and text information. These systems assume descriptions that randomly vary in construction and content, but traditionally assume that description content is homogeneous. This assumption becomes problematic as these models are increasingly extended to descriptions of images of *people* [12], where people are known to show systematic biases in how they process others [18]. Therefore, this paper presents a novel approach to discovering exceptional subgroups of descriptions in which the content of those descriptions reliably differs from the general set of descriptions. We develop a novel interestingness measure for subgroup discovery appropriate for probability distributions across semantic representations and demonstrate its integration into a process for mining exceptional subgroups. The proposed method is applied to a web-based experiment in which 500 raters describe images of 200 people. Our analysis identifies multiple exceptional subgroups and the attributes of the raters and images that produce these descriptions. The implications for intelligent systems based on datasets that contain exceptional descriptions of people are discussed.

1 Introduction

The fields of machine learning and computational linguistics are increasingly focused on building multi-modal intelligent systems that integrate visual and text-based information, e. g., answering questions and generating textual descriptions of images [8, 1, 15], identifying objects in images based on text descriptions [19], or improving the sentence parsing of descriptions by grounding them with visual information [9]. While the data rely on text descriptions of images written by people, people generate idiosyncratic descriptions that differ significantly based on the goals, biases, and expertise of the person writing the description [12].

The issue of detecting heterogeneity of descriptions is particularly important for descriptions of people. People systematically differ in how they encode, search for, and remember faces of other races and genders [18, 13], which may systematically bias the descriptions people generate of others. Importantly, intelligent systems built using descriptions that are assumed to be homogeneous but contain a subset systematically biased data are particularly likely to propagate these implicit biases into the intelligent systems that rely on that data [11, 22].

In this paper we present a novel approach for identifying homogeneous subgroups of descriptions that reliably differ in their content from the general set of descriptions. Specifically, we extract a low-dimensionality representation of individual descriptions based on latent Dirichlet allocation (LDA) [7]. Using that representation, we apply *exceptional model mining* [16, 10], a variant of subgroup discovery [23, 2] that focuses on complex target properties, for detecting homogeneous subgroups of descriptions in the low-dimensional space. Critically, we define a novel quality function based on a subgroups' topic distribution and use it to identify exceptionally unique homogeneous subgroups of textual descriptions.

The efficacy of the proposed approach is validated on a new dataset consisting of 2491 descriptions of 193 faces which has been collected online.[1] We present results of applying the LDA-based exceptional model mining method on that dataset and discuss the implications for intelligent systems based on descriptions generated by people in general, and descriptions of people in particular.
The contribution of the paper is summarized as follows:

1. We present a novel approach for mining exceptional subgroups in descriptions of people using subgroup discovery on topic models using LDA.
2. We introduce a new interestingness measure for subgroups that compares the distribution across topics in subgroups to the overall (expected) distribution.
3. We present and discuss the results of applying the proposed novel methodology to a real-world dataset of descriptions of people collected online.

The remainder of the paper is organized as follows: Section 2 outlines the experimental setup and the dataset. Section 3 describes our novel approach using subgroup discovery on generated topic distributions. Section 4 presents and discusses results. Finally, Section 5 concludes with a summary and presents interesting directions for future work.

2 Experiment

In this section we detail the critical aspects of an online experiment to collect descriptions and judgments about images of people. In the following sections we describe the set of images (i. e., the stimuli) and outline the experimental procedure for obtaining the textual descriptions and ratings of the images.

2.1 Stimuli and Procedure

The stimuli (images) consisted of 193 color images of the head and shoulders of people standing in front of an off-white background. Individuals were instructed to look directly into the camera lens and maintain a neutral facial expression. Multiple images were taken of each individual but only one image was used from each person. These individuals were recruited from the University of Adelaide campus and compensated AUD$10 for their participation.

[1] The experiment was approved by the IRB of the University of Adelaide.

Table 1. Attributes collected for each image description. All values were free response except for ID (random), rater & image genders, as well as typicality and attractiveness.

Self-reported rater attributes	**Rated attributes of the image**	
ID	ID	Eye Color
Age	Age	Hair Color
Gender	Gender	Typicality Rating
Country		Attractiveness Rating

500 participants were recruited via Amazon Mechanical Turk to rate these images and provide written descriptions of the people in the images. Participants were shown five randomly selected images. For each image they were asked to estimate the age, hair color, eye color, ethnicity (selected from a list of African, Asian, Latino, or White), and gender (Male, Female, or Unsure). They were also asked to rate the typicality and attractiveness of the person in the image on a five point scale and guess their occupation. Finally, the participant was asked to write a description of the physical characteristics of the person and a description of the non-physical characteristics the rater felt they could guess about the person (personality, interests, etc.). Each description was required to be at least 10 characters long and consists of at least four words. Participants completing the experiment were paid US$2 for approx. 12 minutes of work.

2.2 Dataset

The experiment resulted in a dataset consisting of 2491 descriptions of 193 faces. Each entry in the dataset contains a set of attributes about the specific rater and the respective image described (Table 1). Unfortunately, the reported ethnicity was incorrectly coded and not recorded. Furthermore, the overall description of each image is given by two free text description elements, one focusing on physical characteristics of the person shown on the image, and the other describing non-physical characteristics. In order to exclude obvious correlations with the physical descriptions/attributes (e. g., gender, attractiveness, eye & hair color, etc.) only the non-physical descriptions were used in all subsequent analyses.

3 Method

The proposed approach consists of the following steps:

1. Given a set of images that are described by a set of *raters* using two sets of textual descriptions, we process the textual descriptions and transform them into a lower dimensional space using topic modeling (LDA).
2. With the resulting topic distributions, we then apply subgroup discovery for mining exceptional subgroups in these descriptions of persons. We obtain the top-k subgroups using an appropriate interestingness measure, e. g., applying our novel quality function comparing the topic Dirichlet distribution of a subgroup to the (expected) overall distribution.

3. Finally, the resulting subgroups are presented for evaluation and assessment with a human-in-the-loop, in order to facilitate interpretation and validation of the obtained results. This is an important step since subgroup discovery is essentially a hypothesis-generating method for identifying exceptional patterns using an (objective) quality function. However, the specific interpretation and interestingness of the pattern must be validated.

3.1 Topic Modeling

The latent Dirichlet allocation model (LDA) [7] is the most popular method of topic modeling in natural language processing. It is a statistical model of how text documents are generated that relies on the assumption that a written text can be represented as a collection of topics. The logic for this representation is that across many documents, only a much smaller number of topics actually exist. Each topic is associated with specific words, and certain words are more or less likely to occur in a document that includes a specific topic. Thus a document composed of words is created by sampling a set of topics, which in turn sample a set of words. Given a set of documents, the LDA model infers the most-likely topics, the words they are associated with, and which documents contain which topics. More formally, the representation for a document d is defined as a probability distribution across topics for that document (θ_d), where the probabilities are drawn from a Dirichlet distribution $\theta_d \sim Dir(\alpha)$. The vector of α values contains an element for each topic, yet a single prior for α is shared for all topics. Furthermore, each topic k is modeled as a probability distribution across words (ϕ_k), where the probabilities are drawn from a Dirichlet distribution $\phi_k \sim Dir(\beta)$. The vector of β values contains an element for each word, yet a single prior value for β is shared for all words in all topics.

Below, we outline the process of constructing a representation for each unique description as a probability distribution across topics. However, the short nature of many individual descriptions prevented the direct application of LDA with each description as a separate document. Instead, the 2491 individual descriptions were combined to create 193 documents, with each document containing all descriptions written based on a single image.

Using this document representation, a grid search across three major parameters of the LDA model (α, β, and the number of topics k) was conducted. The evaluation function for a specific parameterization aimed to maximize the difference between documents while minimizing the number of topics per document. The difference between documents was parameterized as the sum across all pairs of documents of the cosine similarity of probability distribution of topics. The number of topics per document was parameterized as the sum of the conditional entropy across the topic probability distribution for all documents.

Finally, for each of the 2491 individual descriptions a probability distribution across these nine topics was extracted. This probability distribution was used as the target to discover exceptional subgroups in the second analysis phase. For that, we developed a novel quality function which directly assesses the uniqueness of subgroups of topic distributions.

3.2 Subgroup Discovery

For subgroup discovery, intuitively we consider subsets of a dataset (*subgroups*) which are intensionally described by a combination of attribute–value pairs. The interestingness of a subgroup is then determined by an interestingness measure, i. e., a quality function operating on the subgroup and a specific target concept. Using a specific subgroup discovery algorithm, e. g., [5, 3, 17] then the *top-k* patterns are returned.

Formally, a *database* $D = (I, A)$ is given by a set of individuals I and a set of attributes A. For nominal attributes, a *selector* or *basic pattern* $(a_i = v_j)$ is a Boolean function $I \rightarrow \{0, 1\}$ that is true if the value of attribute $a_i \in A$ is equal to v_j for the respective individual. The set of all basic patterns is denoted by Σ. A subgroup is described using a description language, typically consisting of attribute–value pairs. Here, we focus on an exemplary conjunctive pattern description language. A *subgroup description* or (complex) *pattern* P is then given by a set of basic patterns $P = \{sel_1, \ldots, sel_l\}, sel_i \in \Sigma, i = 1, \ldots, l$, which is interpreted as a conjunction, i. e., $P(I) = sel_1 \wedge \ldots \wedge sel_l$, with $length(P) = l$. A *subgroup* $S_P := ext(P) := \{i \in I | P(i) = true\}$,i. e., a *pattern cover* is the set of all individuals that are covered by the subgroup description P. The set of all possible subgroup description, and thus the possible search space is then given by 2^{Σ}. The pattern $P = \emptyset$ covers all instances contained in the database.

A *quality function* $q \colon 2^{\Sigma} \rightarrow \mathbb{R}$ maps every pattern in the search space to a real number that reflects the interestingness of a pattern (or the pattern cover, respectively). For a binary target variable, we can compare the its share in the subgroup to its share in the whole database, typically also weighted by the size of the subgroup, often also using a minimal subgroup size threshold for obtaining meaningful subgroups, cf. [2] for a detailed discussion. For more complex quality functions, we consider functions that take into account a set of target concepts, e. g., [14]. Then, we can compare a set of distributions by utilizing the concept of exceptional models [10].

In the case of topic models, we utilize Dirichlet distributions capturing the overall distribution of topics in the overall dataset (i. e., modeling the expected distribution) while the topic distribution contained in each subgroup is modeled by another Dirichlet distribution. In order to obtain the respective Dirichlet distributions $Dir(\alpha_{\emptyset})$ and $Dir(\alpha_{S_P})$ for the overall population $S_{\emptyset}$ and the subgroup S_P, respectively, we can compute a maximum likelihood estimate (MLE) utilizing the Newton-Raphson method [20, 21] for obtaining the parameter vectors α_{S_P} and $\alpha_{\emptyset}$. For comparing distributions, we utilize the Kullback-Leibler divergence metric *KL*. Thus, for Dirichlet distributions, comparing $Dir(\alpha)$ and $Dir(\beta)$, we obtain

$$\begin{aligned} KL(\alpha, \beta) = \log \Gamma(\alpha_0) - \sum_{i=1}^{K} \log \Gamma(\alpha_k) - \log(\beta_0) \\ + \sum_{i=1}^{K} \log(\beta_k) + \sum_{i=1}^{K} (\alpha_k - \beta_k)(\psi(\alpha_k) - \psi(\alpha_0)) \, , \end{aligned} \tag{1}$$

where Γ is the gamma and ψ is the digamma function. Our novel quality function for comparing topic distributions for a specific subgroup S_P is then given by

$$q_D(P) = KL(\alpha_{S_P}, \alpha_\emptyset)\,, \tag{2}$$

with the distribution parameters α_{S_P} and $\alpha_\emptyset$ for the subgroup and the overall population, respectively.

4 Results and Discussion

After constructing the topic probability distributions for each description, all possible subgroups were evaluated to identify deviating subgroups of descriptions. This was done exhaustively using the SD-Map algorithm [5], provided by the VIKAMINE system [4][2]. We identified deviating subgroups using different values of k for identifying the top-k subgroups, while we discuss results for the top 20 subgroups below; other result sets were consistent. A *minimal improvement filter* [6] was incrementally applied to the set of all subgroups to limit the set of attributes defining exceptional subgroups. Specifically, a specialization P' of a pattern P is considered a more exceptional subgroup if P' improves on the quality function compared to P: So, for example, we consider the specialization of the pattern *face_hair_color* = *black* to *face_hair_color* = *black AND face_gender* = *male*, if the quality of the latter pattern increases. A minimal subgroup size threshold of 1% was used in this analysis.

In the analysis presented below, we utilized the proposed quality function $q_D(P)$. These results were compared with the standard Hotelling quality function [2], which did not produce as coherent subgroup attributes as the novel quality function $q_D(P)$. This is likely to be due to the direct correspondence between the new quality function $q_D(P)$ and the multinomial probability distribution that comprises the representation of a description across topics.

First, in Section 4.1 we present an analysis of the types of attributes more likely to define deviating subgroups and the relationship between attributes. Second, in Section 4.2 we aggregate across exceptional subgroups and evaluate the frequency with which specific images are the subject of descriptions in exceptional subgroups and that specific raters generate exceptional subgroups. Finally, we discuss the importance of applying these results for intelligent systems that rely on potentially heterogeneous descriptions generated by people.

4.1 Attributes

Interesting patterns emerge across the seven attributes that identify the 20 subgroups most dissimilar to the overall set of descriptions, see Table 2 for a detailed view. Though the gender of the rater and the face in the image were two of the most common attributes to define exceptional subgroups (in our result set), only

[2] http://www.vikamine.org

Table 2. Exceptional subgroup attribute frequencies. The attributes and values of a description that are indicative of the top 20 exceptional subgroups. The count column indicates the number of subgroups that are distinguished by this attribute. R and I in the column headers indicate Rater and Image attributes, respectively.

R. Country	R. Gender	I. Gender	I. Eye Color	I. Hair Color	I. Ratings
USA (3)	Female (3)	Female (8)	Black (12)	Black (6)	Typicality (4)
India (2)	Male (7)	Male (3)	Brown (1)	Blond (1)	Attract. (5)
			Green (1)		

four subgroups were defined by both the gender of the rater and face. This suggests the interaction between rater and face gender was no more likely to identify an exceptional subgroup than either factor independently.

A similar pattern emerges for the typicality and attractiveness ratings of the faces. The two subjective attributes of the faces were defining attributes for 8 of the 20 subgroups, but only co-occurred as defining characteristics in one subgroup (a group of 20 descriptions consisting of images of females where typicality was rated "agree" and attractiveness was rated "strongly-agree").

The eye color of the image was a defining attribute of 14 of 20 exceptional subgroups and the attribute value "black" was by far the most frequently occurring value. This is particularly surprising given that eye color attribute was described as "black" for only 13% of descriptions, which was second frequent after "brown" and more frequent than "blue" and "green." One possible explanation of this pattern is that perceived eye color is highly correlated with perceived ethnicity or race for these descriptions, a possibility we discuss further in the discussion section.

4.2 Image and Rater Distributions

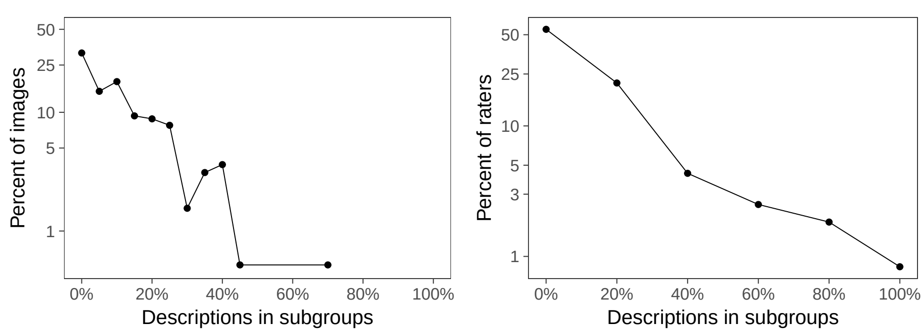

Fig. 1. The proportion of descriptions of specific images (left panel) and raters (right panel) that occur in at least one exceptional subgroup.

Given that 736 of 2491 descriptions were assigned to one of the 20 most distinct subgroups, it is perhaps not overwhelmingly surprising that only 31% of images did not appear in any subgroup (see left panel of Figure 1). Despite most images occurring in at least one subgroup, only nine images had at least 40% of their descriptions included in a subgroup and only one image had more than 50%.

A different frequency pattern emerges for raters, as shown on the right of Figure 1. Unlike images, where at least one description was likely to be in a deviating subgroup, more than 50% of raters do not have a description in any deviating subgroup. Furthermore, the majority of raters who did produce a description that occurred in a deviating subgroup produced only one such description.

Despite the high number of raters with one or zero descriptions in the exceptional subgroups, Figure 1 shows a stronger positive skew for raters than images: a small but significant group of raters, 5.2% of the sample, had more than 50% of their descriptions be identified as belonging to at least one deviating subgroup.

4.3 Discussion

These results suggest that descriptions of people that significantly deviate from the population of descriptions are relatively frequent. Furthermore, these exceptional descriptions are not exclusively driven by particularly exceptional images or particularly exceptional raters. Instead, the vast majority of descriptions that are identified as exceptional are descriptions from raters for whom most descriptions are not exceptional, and of images whose descriptions are mostly not exceptional.

The attributes that define the maximally deviating subgroups point to the types of features of raters and images that are likely to produce exceptional descriptions. Male raters and female images are attributes that are likely to define deviating subgroups, though these two attributes appear to independently contribute and not in combination. Additionally, in this population of images, black hair and black eyes are the attributes of images most likely to identify exceptional subgroups. Further work is necessary to understand if these attributes produce exceptional descriptions when embedded in a different sample of images, or if these attributes are predictive of latent attributes, such as ethnicity, that were not included as attributes for the subgroup discovery process.

Taken as a whole, these results do not show strong evidence of subgroups of descriptions that are identified based on gender, age, or race [18, 13], perhaps decreasing the fear that intelligent systems based on descriptions of people will inherit strong implicit biases from the raters [11]. However, we do find certain attributes of images, particularly black eye color and black hair color, which are much more likely to produce exceptional descriptions than other attributes. This suggests that the topics and words people use when describing the non-physical characteristics of other people may vary widely. The degree to which the importance of these attributes are an artifact of the particular faces we studied is an open question, but it highlights the need for further evaluation of these datasets.

These issues are increasingly important as intelligent systems, trained with labels and descriptions generated by people, become ubiquitous. These systems rely on human annotated descriptions, and people are clearly not homogeneous in how they view other people. When combined with methods like LDA for extracting semantic representations of descriptions, exceptional model mining and subgroup discovery techniques can provide a necessary tool to help identify potential biases in these descriptions. Additionally, these tools can possibly suggest specific images, subgroups, and attributes where additional data would help alleviate the bias in the systems that rely on them.

5 Conclusions

This paper presents a novel method of combining topic modeling and subgroup discovery to identify interesting image descriptions. Critically, we present a novel definition of interestingness that compares the subgroup and general population using the Kullback-Leibler divergence between the Dirichlet distribution that characterizes the probability distribution of topics. This method is applied to the problem of subgroup discovery among descriptions of pictures of people, a domain that has broad implications for applied domains [12] while carrying a real risk of biased descriptions [18].

Our analysis method detects meaningful subgroups of image descriptions that diverge from the general set of descriptions and characterizes them based on both properties of the raters as well as the images. These subgroups suggests new norms for data collection methods and statistical models for web-based applications that are sensitive to the heterogeneous nature of descriptions of people.

For future work, we aim to extend the analysis and data collection in order to investigate (dis-)similarities in more datasets and applications. In particular, we aim to investigate specific characteristics with respect to the subgroup profiles and their distributions, also including novel inspection and validation approaches. Furthermore, the inclusion of contextual domain knowledge into the models is an interesting issue to consider.

6 Acknowledgments

Funding for data collection was provided by a University of Adelaide Interdisciplinary Research Grant to C. Semmler, A. Hendrickson, R Heyer, A. Dick and A. van den Hengel. Furthermore, this work has also been partially supported by the German Research Council (DFG) under grant AT 88/4-1.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual Question Answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)

2. Atzmueller, M.: Subgroup Discovery – Advanced Review. WIREs Data Mining and Knowledge Discovery **5**(1), 35–49 (2015). https://doi.org/10.1002/widm.1144
3. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. ISMIS. LNCS, vol. 5722, pp. 1–15. Springer (2009)
4. Atzmueller, M., Lemmerich, F.: VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: Proc. ECML/PKDD (2012)
5. Atzmueller, M., Puppe, F.: SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In: Proc. PKDD. pp. 6–17. Springer, Heidelberg, Germany (2006)
6. Bayardo, R., Agrawal, R., Gunopulos, D.: Constraint-Based Rule Mining in Large, Dense Databases. Data Mining and Knowledge Discovery **4**, 217–240 (2000)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
8. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient Object Detection: A Benchmark. IEEE Transactions on Image Processing **24**(12), 5706–5722 (2015)
9. Chrupała, G., Gelderloos, L., Alishahi, A.: Representations of Language in a Model of Visually Grounded Speech Signal. In: Proc. ACL. pp. 613–622. Association for Computational Linguistics (2017)
10. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional Model Mining. Data Mining and Knowledge Discovery **30**(1), 47–98 (2016)
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. CoRR **abs/1104.3913** (2011)
12. Gatt, A., Tanti, M., Muscat, A., Paggio, P., Farrugia, R., Borg, C., Camilleri, K., Rosner, M., van der Plas, L.: Face2Text: Collecting an Annotated Image Description Corpus for the Generation of Rich Face Descriptions. In: Proc. 11th edition of the Language Resources and Evaluation Conference (LREC'18) (2018)
13. Herlitz, A., Lovén, J.: Sex Differences and the Own-Gender Bias in Face Recognition: A Meta-Analytic Review. Visual Cognition **21**(9-10), 1306–1336 (2013)
14. Klösgen, W.: Handbook of Data Mining and Knowledge Discovery, chap. 16.3: Subgroup Discovery. Oxford University Press, New York (2002)
15. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
16. Leman, D., Feelders, A., Knobbe, A.: Exceptional Model Mining. In: Proc. ECML/PKDD. LNCS, vol. 5212, pp. 1–16. Springer (2008)
17. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. Data Mining and Knowledge Discovery **30**, 711–762 (2016). https://doi.org/10.1007/s10618-015-0436-8
18. Levin, D.T.: Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit. Journal of Experimental Psychology: General **129**(4), 559–574hypo (2000)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
20. Minka, T.: Estimating a Dirichlet distribution (2000)
21. Sklar, M.: Fast MLE Computation for the Dirichlet Multinomial. arXiv preprint arXiv:1405.0099 (2014)
22. Torralba, A., Efros, A.A.: Unbiased Look at Dataset Bias. In: Proc. EEE Conference on Computer Vision and Pattern Recognition. pp. 1521–1528. IEEE (2011)
23. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. PKDD. pp. 78–87. Springer, Heidelberg, Germany (1997)