

Constrained Latent Dirichlet Allocation for Subgroup Discovery with Topic Rules

Rui Li¹ and Zahra Ahmadi² and Stefan Kramer³

Abstract. Subgroup discovery is the task of identifying subgroups that show the most unusual statistical (distributional) characteristics with respect to a given target variable, at the intersection of predictive and descriptive induction. Redundancy and lack of rule interpretability constitute the major challenges in subgroup discovery today. We address these two issues by constrained latent Dirichlet allocation (LDA) to identify co-occurring feature values (descriptions) for subgroup rule search, obtaining a less redundant and more diverse rule set. Latent Dirichlet Allocation, as a topic modeling approach, is able to identify diverse topics, from which the rules can be derived. The resulting rules are less redundant and can also be interpreted by the corresponding topic. Experimental results on six benchmark datasets show that the presented approach provides rule sets with better rule redundancy and diversity compared to those of four existing algorithms. One unique and interesting advantage of the proposed method is that it can categorize rules by topics as well as the assignment of a probability to each feature value of a discovered rule, which can be used in the interpretation of the results.

1 Introduction

Subgroup discovery (SD) aims at identifying subgroups described by conjunctions of feature values that are statistically most interesting with respect to a given target variable [14, 28]. It is a task at the intersection of predictive and descriptive induction. For example, a subgroup rule may be “if house = own and job = skilled, then credit rating = good”, where “house” and “job” are the features (attributes), “own” and “skilled” are their corresponding feature values or feature conditions, and “credit rating = good” is the target variable. Subgroup rules have conjunctions of feature values on the left-hand side and a user-specified target class on the right-hand side.

There are several important issues concerning subgroup discovery. First of all, the search strategy is an intensively studied topic, because the search space grows exponentially as the dimension increases. Thus, investigating all of the possible feature value combinations is simply infeasible for high dimensional data. To cope with it, beam search is used to explore only a tractable fraction of the search space. On the other hand, the optimistic estimate [10, 28] is another alternative that discards the non-promising search branches and only concentrates on the top most promising subgroups at each level. The second essential aspect is the level of redundancy. During the process of subgroup rule mining, many similar rules can be found, al-

though they all pass the selection criterion (e.g., quality measure). However, they may be some variants of the same scheme. Thus, discovering qualified but also redundancy reduced (diversity increased) subgroups is of great interest [17, 24]. Also, too many rules make it hard for users to interpret and validate the results. To address these issues of subgroup redundancy and interpretability, we approach the problem of subgroup discovery from a statistical perspective.

Motivated by the goal of rule interpretability, we conjecture it is easier to interpret rules once they are categorized, because categorization can reveal similarity/dissimilarity. In documents categorization, latent Dirichlet allocation (LDA) [4] is a generative topic modeling approach to identifying co-occurring words in documents. Each document can be characterized by a set of topics, and each topic is associated with a set of words. The popularity of LDA and its extension spreads across different application areas, such as document clustering, routine discovery, and so forth.

Subgroup discovery aims at finding conjunctions (co-occurrences) of feature values that together predict a target. On the other hand, LDA is meant to find co-occurring words in documents. Hence, both techniques uncover co-occurring patterns (words in LDA and feature values in SD). Thus, it is then feasible to lend the idea of LDA to SD to effectively find rules, without exhaustively searching the prohibitively large space of rules. Besides, a recent study [11] has shown that the use of the Dirichlet process [23], closely related to LDA, is efficient in finding frequent itemsets in binary transaction data. In addition, an Entity Topic Model (ETM) approach [13] was presented to devise topic models for documents with entity information by capturing the word co-occurrences. Inspired by this work, we present a constrained latent Dirichlet allocation (CLDA) approach to SD. Its main contributions are as follows:

- It offers another way of integrating LDA into SD to find interesting rules (a related method was proposed by Atzmüller and Mitzlaff [3]).
- A tailored CLDA is proposed to practically bring LDA and SD together.
- The resulting rules can be interpreted and categorized by various discovered topics, which is missing in existing SD algorithms.

The rest of the paper is organized as follows. In Section 2, related work regarding SD and redundancy management is reviewed. Section 3 proposes the CLDA for subgroup discovery, followed by Section 4 with experimental results. Some conclusions are drawn in Section 5.

2 Related Work

As a local pattern mining methodology, subgroup discovery (SD) is closely related to other techniques. For example, emerging pat-

¹ Informatik/I12, Technische Universität München, Germany, email: rui.li@in.tum.de

² Informatik, Johannes Gutenberg - Universität Mainz, Germany, email: zahmadi@uni-mainz.de

³ Informatik, Johannes Gutenberg - Universität Mainz, email: kramer@informatik.uni-mainz.de

tern mining, class-correlated pattern mining and contrast set mining [20]. Recently, an extended framework [21] was developed to handle constraint-based sequential pattern mining, using a sequential pattern growth methodology. Regarding research on SD, the pioneering work EXPLORA [14] presented a single-relational setting SD method. A multi-relational setting, MIDOS [28], used the concept of optimistic estimates to ensure efficiency. This work also introduced sampling to estimate the relative frequency of a group via the Chernoff bound. Another study [22] proposed a knowledge-based sampling method, offering a generic way of incorporating prior knowledge in the sense of sampling. CN2-SD [15] employed the weighted relative accuracy measure to trade generality against accuracy, as well as re-weighted the samples in rule discovery. A tight optimistic estimate [10] was proposed to limit the search space size. The tight optimistic estimate (TOE) is a tight upper bound of the quality of the promising subgroups. Similarly, Difference-based Estimates for a Generalization-Aware (DBEGA) [16] method was presented to take generalization into account using some new optimistic estimates. More recently, an interactive subgroup discovery approach was proposed to allow user feedback during search, obtaining more interesting and diverse rules [6].

Redundancy is an essential issue in pattern mining, which has been particularly actively studied in recent years. The MDL (minimum description length) principle has witnessed a renaissance in this regard. The KRIMP algorithm [26] uses an itemset coding scheme to compress the data best. Related work employed a probabilistic maximum entropy model to iteratively find itemsets [18]. The authors suggested the use of MDL to identify itemsets that summarize the data well. A closure system [5] was suggested to search for extensions of quality functions rather than their individual descriptions. The authors found that the search space and outputs were efficiently reduced when equivalence classes are considered. A relevant study Diverse Subgroup Set Discovery (DSSD) [25] attempted to obtain diverse rule sets by integrating a diverse beam search into each level-wise search. Subgroup Discovery via Quadratic Programming (SD-VQP) [17] suggested to efficiently attain less redundant rules by selecting interesting features using optimization. A random sampling method to obtain maximal itemsets was proposed, which can make use of any monotonically decreasing measure as interestingness criterion [19].

3 Constrained Latent Dirichlet Allocation (CLDA) for Subgroup Discovery

He and Shapiro [11] attempted to discover frequent itemsets in binary data using the Dirichlet process. Our proposed algorithm CLDA is not confined to binary data, and we are also able to identify topics indicating subgroup rules. We illustrate how to incorporate the idea of LDA into SD to find interesting subgroup rules. A short introduction to LDA is given first.

LDA is a generative model, describing how observed words in documents can be generated by hidden (latent) topics. The structure is shown in Fig. 1. In this model, there is a set of N words $\mathbf{w} = \{w_1, \dots, w_N\}$, M documents $D = \{d_1, \dots, d_M\}$, with each w_i belonging to some document d . In each document, there is a multinomial distribution over K topics, which are not directly observable (latent). Each document is associated with a parameter θ^d . A topic is also modeled as a multinomial distribution over words. Two hyperparameters α and β are introduced in this model. A low hyperparameter α encourages few topics per document and low β encourages few words per topic.

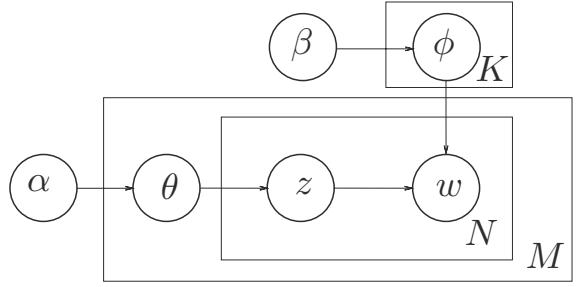


Figure 1. Graphical model of smoothed latent Dirichlet allocation [4]. θ is the per-document topic assignment and z is the per-word topic assignment. ϕ is the per-corpus topic distribution parameterized by β .

The LDA generative process can be explained as follows: Draw M multinomials θ from a Dirichlet prior α , one for each document; draw K multinomials ϕ from a Dirichlet prior β , one for each topic; draw a topic z_i from multinomial θ , i.e., $p(z_i|\alpha)$ and then draw a word w_i from multinomial ϕ , i.e., $p(w_i|z_i, \beta)$.

$$\theta \sim \text{Dirichlet}(\alpha) \quad (1)$$

$$z_i|\theta^d \sim \text{Multinomial}(\theta^d) \quad (2)$$

$$\phi \sim \text{Dirichlet}(\beta) \quad (3)$$

$$w_i|z_i, \phi \sim \text{Multinomial}(\phi_{z_i}) \quad (4)$$

The central task of inference in LDA is to determine the posterior distribution of latent topic variables θ and z conditioned on the words in the documents. Apart from the mean field variational methods [4], a collapsed Gibbs sampling approach, one kind of Markov chain Monte Carlo, was proposed to yield the solution [9]. The ratio of probability of w_i under topic j is:

$$p(w_i|z_i = j, z_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(.)} + W\beta}, \quad (5)$$

where $n_{-i,j}^{w_i}$ is the number of times a word assigned to topic j , excluding the current one, and $n_{-i,j}^{(.)}$ is the total number of times of words assigned to topic j , excluding the current one. W is the number of distinct words. The probability of topic j in document d is:

$$p(z_i = j|z_{-i}) = \frac{n_{-i,j}^d + \alpha}{n_{-i}^d + K\alpha}, \quad (6)$$

where $n_{-i,j}^d$ is the number of times a word from document d assigned to topic j , not including the current one, and n_{-i}^d is the total number of times a word in document d , excluding the current one. Thus the full conditional posterior distribution for z_i ($p(z_i = j|z_{-i}, \mathbf{w})$) can be obtained as:

$$\begin{aligned} p(z_i = j|z_{-i}, \mathbf{w}) &= \frac{p(z_i = j, w_i|z_{-i}, \mathbf{w}_{-i})}{p(w_i)} \\ &\propto p(z_i = j, w_i|z_{-i}, \mathbf{w}_{-i}) \\ &= \underbrace{p(w_i|z_i = j, z_{-i}, \mathbf{w}_{-i})}_{\text{likelihood}} \cdot \underbrace{p(z_i = j|z_{-i})}_{\text{prior}} \\ &= \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(.)} + W\beta} \cdot \frac{n_{-i,j}^d + \alpha}{n_{-i}^d + K\alpha}. \end{aligned} \quad (7)$$

class	feature 1	feature 2		T1	T2	T3	T4
+	A_1	D_2		A_1	0	2	0
+	C_1	D_2	→	B_1	0	0	4
+	B_1	D_2		C_1	0	0	5
-	C_1	E_2					30
⋮	⋮			$p_{A_1,1}^1 = 1$ (T1)	$p_{A_1,3}^1 = \frac{1}{4+5+3}$ (T3)		
				$p_{A_1,2}^1 = 1$ (T2)	$p_{A_1,4}^1 = \frac{20}{10+30+20}$ (T4)		

Figure 2. Numerical demonstration of CLDA regarding the four cases in Eq. 9. Feature 1 constitutes three distinct values (A_1, B_1, C_1), in which A_1 is used to show the calculation of $p(w_{i,j}^k)$ assuming four topics T1 to T4. Note the samples in the positive and the negative class are used separately.

Bringing SD and LDA together via CLDA: In the proposed method, a “feature=value” expression functions as a word in the topic model, thus the number of total distinct feature values in the data amounts to the total number of words in LDA. We also assume that there are topics expressing some perspectives on the data. Thus, the subgroup rules can be immediately discovered after inferring the topics and their associated feature values. Feature values from the same feature may be grouped into the same topic, whereas the rule conditions in SD should be from different features. Therefore, we should effectively impose some constraints that encourage feature values from the identical feature to go into different topics. To this end, we propose a CLDA approach tailored for finding subgroup rules.

Recently, CLDA [29, 2] was suggested to allow the use of prior knowledge. The cannot-link and must-link constraints were realized by incorporating a term in Eq. 7. It can be shown [1] that the conditional probability can be altered by multiplying a factor at the right-hand side of Eq. 7. Differing from their work, we suggest a different form of the constraint devised for SD. For example, we can intentionally multiply it with 0 if we knew a word belonging a topic j with probability 0. Similar to the conducted work [29, 2], we allow a soft constraint modifying Eq. 7 as:

$$p(z_i = j | z_{-i}, \mathbf{w}) = p(w_{i,j}^k) \cdot \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^d + \alpha}{n_{-i}^d + K\alpha}, \quad (8)$$

where $p(w_{i,j}^k)$ denotes the prior probability of a feature value w_i from feature k belonging to topic j , and it is computed as:

$$p(w_{i,j}^k) = \begin{cases} 1 & \text{if } n_{-i,j}^k = 0, n_{i,j}^k = 0 \\ 1 & \text{if } n_{-i,j}^k = 0, n_{i,j}^k \neq 0 \\ \frac{1}{n_{-i,j}^k + n^k} & \text{if } n_{-i,j}^k \neq 0, n_{i,j}^k = 0 \\ \frac{n_{-i,j}^k}{n_{-i,j}^k + n_{i,j}^k} & \text{if } n_{-i,j}^k \neq 0, n_{i,j}^k \neq 0 \end{cases} \quad (9)$$

where $n_{i,j}^k$ is the number of times of feature value w_i from feature k belonging to topic j . $n_{-i,j}^k$ is the number of times this topic j already assigned to the feature k excluding the current w_i . n^k is the number of distinct feature values in feature k , which is used to act as a Laplace smoothing term. The essence of $p(w_{i,j}^k)$ is to encourage feature values from the same feature to fall into different topics by investigating the previous topic assignments. Fig. 2 demonstrates that calculation of $p_{A_1,j}^1$ is only involved with prior counting statistics of A_1, B_1 and C_1 , regardless of D_2 from the second feature. As for topic 1 (T1) and 2, no prior statistics of other feature values B_1 and

Algorithm 1 CLDA for Subgroup Discovery with Topic Rules

Input: K : allowed maximal number of topics, training data $\mathcal{D}_{\text{train}}$, test data $\mathcal{D}_{\text{test}}$
Output: Collected rules SR (subgroup rules)

- 1: Data preparation for CLDA
- 2: **for** $i = 1$ to K **do**
- 3: Run CLDA on positive and negative samples from $\mathcal{D}_{\text{train}}$, respectively
- 4: Calculate the perplexity using Eq. 10 based on $\mathcal{D}_{\text{test}}$
- 5: **end for**
- 6: Determine an appropriate number of topics K_{best} based on calculated perplexity
- 7: **for** $j = 1$ to K_{best} **do**
- 8: Choose the corresponding features of co-occurring feature values produced from positive and negative samples respectively, as candidates to find SR on the training data $\mathcal{D}_{\text{train}}$
- 9: Collect the rules SR
- 10: **end for**

C_1 is given, therefore the prior probability belonging to the topic is 1. As for topic 3, B_1 and C_1 are already assigned to it with 4 and 5 times. Thus, its prior probability for this topic is only $\frac{1}{4+5+3}$, where 3 is the number of distinct feature values in feature 1, i.e., A_1, B_1 and C_1 . In terms of topic 4, the probability is proportional to its assignment 20 over the total assignments 60.

SD needs a class label (supervised) to find the rules, whereas LDA (unsupervised) does not request any class information. Thus, we divide the data into positives and negatives, constructing the CLDA based on data from each of the two classes, respectively. When built on either of the classes, CLDA produces the co-occurring feature values regarding the respective target class. It is equivalent to state that feature values tend to appear together in the positive or negative class, which is in line with the goal of finding rules pointing to a given target. In Alg. 1, line 7 to line 10 are devoted to finding the actual SD rules with a fixed number of topics K_{best} . For each topic, we have some feature values associated with integers indicating the number of assignments. The larger the number, the more frequently it appears in that topic, and of course zero means no occurrence. We then find the *corresponding features* of these *feature values* for exhaustive SD rule search using the quality function of Eq. 12. One can also only examine the combinations of these feature values for SD rules, but this may limit the number of discovered rules. In particular, we suppose that some features as a whole describe a certain topic, therefore we execute the search in a broader space.

Data Preparation for CLDA: Line 1 in Alg. 1 prepares the data for running CLDA. If the data is numeric, then we first discretize them into nominal. The data may be denoted as integers, such as 1, 2, etc. Thus, two different features can have the same feature value of, for example, 1, but 1 in a feature is different from 1 in another feature. We, hence, intentionally denote each feature value uniquely to form a set of feature values (just as a vocabulary in documents). As a result, each sample is represented by some feature values drawn from the feature value set. In Fig. 2, for example, the set is $A_1, B_1, C_1, D_2, \dots$ for the positive class and C_1, E_2, \dots for the negative class.

Choosing the Number of Topics: It is often hard to know the number of topics in advance. One common remedy known from language modelling is the use of per-word predictive perplexity (low values are suggested) as a measure of the likelihood of the model based on a held-out test set [4]. It is a measure of the generalization

ability of the model on unseen data. Theoretically, one can choose the best number of topics according to the lowest perplexity. We applied the perplexity suggested by Heinrich [12], which can be briefly formulated as:

$$\text{perplexity}(\mathbf{w}_{\text{test}}) = \exp \left\{ -\frac{\sum_{d=1}^{\mathcal{D}_{\text{test}}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{\mathcal{D}_{\text{test}}} N_d} \right\}, \quad (10)$$

where $\sum_{d=1}^{\mathcal{D}_{\text{test}}} N_d$ is the total number of feature values in the test data. $p(\mathbf{w}_d)$ is calculated as:

$$p(\mathbf{w}_d) = \prod_{i=1}^N \left(\sum_{j=1}^K \phi_{j,i} \cdot \theta_{d,j} \right)^{n_i^d}, \quad (11)$$

where n_i^d is the number of times feature value (word) i appears in test sample (document) d . In this study, $n_i^d = 1$ because a feature can appear only once in a sample. $\phi_{j,i}$ is calculated using Eq. 5 only from training data. $\theta_{d,j} = \frac{n_j^d + \alpha}{\sum_{j=1}^K n_j^d + K \cdot \alpha}$, where n_j^d is the number of times a feature value assigned to topic j calculated using Eq. 6. For details, see Eq. 93 of the original reference [12].

Subgroup Discovery: Quality Function: Subgroups are usually evaluated by a quality function providing a trade-off between rule generality and distributional unusualness. Perhaps the most common form is:

$$\varrho = g^a(p - p_0), \quad (12)$$

where p is the rule accuracy (support), i.e., the fraction of rows of the target class in the subgroup, and p_0 is the default rule accuracy, i.e., the fraction of rows of the target class in the database. g is the generality (coverage)⁴ of the subgroup. Parameter a , between 0 and 1, controls the effect of accuracy by weighting the generality. $a = 1$ or 0.5 is commonly used, and $a = 1$ is used in the present work.

Evaluation Measures: Subgroup discovery can be evaluated by different measures [15], we focus on the following four measures for our comparison.

Cover Redundancy (CR) [17, 24]: It measures the cover count of each sample covered by the rule set, and the deviation from the mean cover count is used to judge the level of redundancy. If the rule set covers some samples unevenly and ignores the others, then this rule set focuses too much on one part of the data. Hence, it probably has some degree of redundancy. Therefore, a lower CR suggests the subgroup rule set covers the data fairly well and is less redundant. Denote a dataset as \mathcal{D} and a set of subgroups \mathcal{S} . The cover count (CC) of a sample m is simply how many times this sample is covered by the rule set \mathcal{S} , i.e., $CC(m, \mathcal{S}) = \sum_{s \in \mathcal{S}} \mathcal{D}_s(m)$. The expected count $\overline{CC} = \frac{1}{|\mathcal{D}|} \sum_{m \in \mathcal{D}} CC(m, \mathcal{S})$. The CR is then computed as:

$$CR^{\mathcal{D}}(\mathcal{S}) = \frac{1}{|\mathcal{D}|} \sum_{m \in \mathcal{D}} \frac{|CC(m, \mathcal{S}) - \overline{CC}|}{\overline{CC}}. \quad (13)$$

The CR is supposed to compare different subgroup sets of (roughly) the same size for the same dataset [24].

Jaccard Index (JI): It is employed as a measure of the diversity of a rule set. Given rules r_1 and r_2 from a rule set \mathcal{R} , the JI is calculated as:

$$JI(r_1, r_2) = \frac{|r_1 \cap r_2|}{|r_1 \cup r_2|}. \quad (14)$$

⁴ Note that coverage is often called support (frequency) in the itemset mining literature, whereas support is in fact accuracy in SD. We follow the convention from the SD literature.

The rules have common elements only when they have matched feature values. As JI (the lower, the more diverse) is computed in a pair-wise manner, we compute it for every two rules in the rule set \mathcal{R} . Then the mean JI is $\frac{\sum_{i=1}^n JI_i}{n}$, where $n = \binom{|\mathcal{R}|}{2}$ is the number of comparisons.

Accuracy reflects the predictive power of the resulting rule set.

Number of Rules is related to the amount of time a human may need to examine and interpret the rules.

Insights into Rules by Topics: The four evaluation measures allow a comparison among SD algorithms, while one merit of the proposed approach is that it offers the possibility of getting deeper understanding of rules by investigating the topics. To gain further insights into the rules in various topics, we suggest a measure of pair-wise distance between every two set of rules in two topics. The distance is measured on every pair of rules in the two topics. The rules in a topic are called a topic rule set. We define a rule distance rd for single rules r_1 and r_2 as:

$$rd(r_1, r_2) = \frac{\text{Hamming distance}(r_1, r_2)}{\max(|r_1|, |r_2|)}. \quad (15)$$

The Hamming distance measures the bitwise difference between two rules, and the denominator ensures the measure is bounded by $[0, 1]$. It measures the dissimilarity/distance, as opposed to JI , a measure of similarity. For example, $rd(\{A\}, \{B, C\}) = 2/2$ and $rd(\{A\}, \{A, B, C\}) = 2/3$. The calculated pair-wise distance can be used to show a dendrogram as Fig. 5.

Compared Methods: Four methods (cf. Fig. 4) are employed to compare with the proposed CLDA. These methods were already introduced in Section 2. They represent a diverse set of methods regarding SD, e.g., optimistic estimate, redundancy reduction and diversity. In DSSD, default parameters were used, except minCoverage = 1 and maxDepth = 4. We only chose the “equal” rule descriptions to stay the same with other methods. Certainly, there are many other SD algorithms can be compared with, but the chosen ones are the most recent approaches.

Table 1. Description of six UCI datasets [8]. †: samples with missing values were removed. Att.: attributes. ‡: multi-class datasets were converted to binary by merging several classes into one, i.e., the largest versus the rest. The continuous attributes were discretized by entropy-based discretization.

Dataset	#Samples	#Att.	#Classes
pima	768	8	2
glass‡	214	9	6
heart statlog [heart]	270	13	2
credit approval [credit]†	653	15	2
bank	4521	17	2
german credit [GC]	1000	20	2

4 Experiments

The algorithms are tested on six UCI datasets [8]. A 10-fold cross-validation was conducted to hold-out some data for calculating perplexity. The tested number of topics (T) was from 5 to 100. The hyperparameters were set to $\beta = 0.1$ and $\alpha = 50/T$ (same as in previous work [7, 9, 27]), where T is the number of topics (i.e., the testing number i in Alg. 1). These values of hyperparameters turned out to be suitable also in our tests. For the CLDA inference, we implemented a collapsed Gibbs sampling approach with 500 iterations.

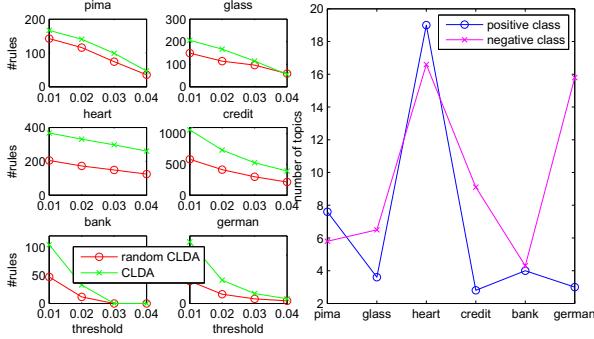
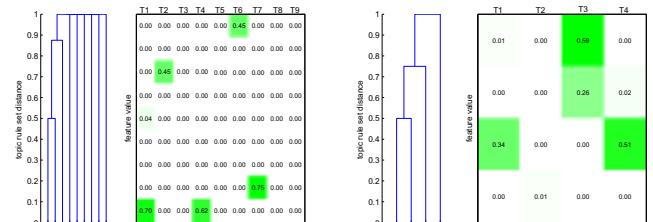


Figure 3. Left figure shows the relation between number of rules and thresholds using random CLDA and CLDA methods. Right figure is the discovered actual number of topics on the six UCI datasets. The numbers are averaged over positive and negative classes from 10-fold cross-validation.



(a) pima

(b) glass

(c) heart

(d) credit

(e) bank

(f) german

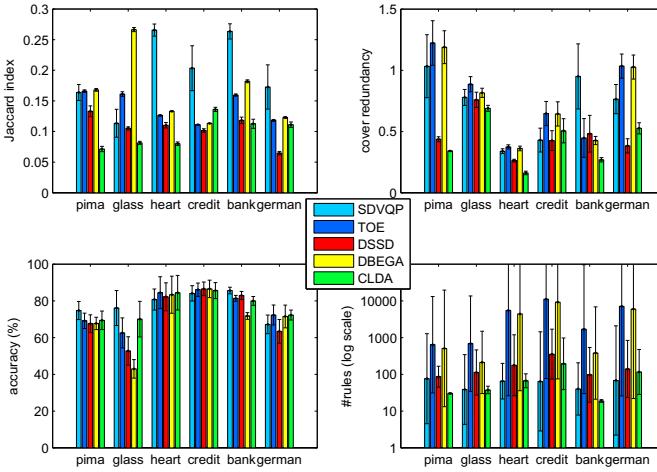


Figure 4. Evaluation measures of methods. The error bar represents the standard deviation from 10-fold cross-validation.

The threshold was $\delta = 0.01$ for the SD quality function. The rules were post-processed by the likelihood-ratio χ^2 test [15] at a significance level of 0.05.

4.1 Comparison with Baseline Random CLDA

We first empirically show that the proposed CLDA is feasible by comparing it with randomly chosen features, i.e., random CLDA. For each topic in CLDA, the algorithm suggests some feature values co-occurring often, whose respective features are then used to identify the actual SD rules. Instead of using these identified co-occurring feature values, we *randomly* selected the same amount of feature values for rule search. Fig. 3 clearly shows that CLDA yields many more rules than random CLDA, which proves that CLDA can indeed find co-occurring feature values.

Fig. 3 illustrates that the number of topics is not influenced by number of samples and dimensions. Regarding the positive class, heart reveals 19 topics, while german has only three topics despite it has the most samples and dimensions. The number of topics in the negative class varies slightly across these datasets.

Figure 5. Dendrogram and calculated probability matrix (cf. Eq. 5) of feature values associated with yielded topics of the positive target class. Note that the sum of feature value probability in a topic is not one because the matrix shows only qualified SD rules.

4.2 Results on Six Datasets: Evaluation Measures

In terms of the Jaccard index (JJ) measure, CLDA shows the lowest value on the pima, glass, heart and bank data. DSSD indicates a lower value on the remaining two datasets (credit and german), as it is particularly devised to discover *diverse* rules. As for redundancy, CLDA holds again the lowest value on pima, glass, heart and bank, being slightly worse (higher value) than DSSD on credit and german. The other three methods exhibit greater values than CLDA and DSSD overall. In terms of accuracy, all these methods show similar results, with SDVQP performing three times the best, on pima, glass and bank. The reason is that SDVQP integrates mutual information between feature and target class into the process of uncovering SD rules, therefore it has good predictive power. Regarding the number of discovered rules, TOE has the greatest number, since it finds all the qualified rules by shrinking the search space via an optimistic estimate [10]. Next to TOE, DBEGA also returns more rules than the other three methods because it is a similar approach as TOE but focusing on generalization aware SD rules. SDVQP, DSSD and CLDA are not devised to find *all* SD rules, they are rather aiming for diverse rule sets. Thus, the size of the resulting rule set is smaller. In summary, CLDA returns rule sets within the same accuracy range as the other methods, but with comparatively low redundancy.

4.3 Results on Six Datasets: Insights into Rules by Topics

By design, CLDA also facilitates easier rule interpretation by categorizing rules into various topics. The dendograms in Fig. 5 show that topics can be grouped by measuring their rules' similarity. If there are many rules, we can interpret and examine these rules by looking at their topics. Choosing topics far apart in the dendrogram gives quite dissimilar rules, and choosing topics near each other gives similar rules. Hence, it is possible to interpret the SD rules via *uncovered hidden* topics. Take the glass dataset for example: Topic three (T3) and four (T4) are neighbors by sharing the second feature value marked by light green. In addition, CLDA gives a probability assignment to each of the feature values in every topic. This probability reveals how likely this feature value belongs to the topic.

5 Conclusions

This paper presented a constrained latent Dirichlet allocation (CLDA) approach to discovering less redundant and more diverse subgroup rules. Instead of exhaustively searching the space of rules, we use a topic modeling method CLDA to identify co-occurring feature values. The feature values are associated with hidden topics, which are uncovered and used to find the actual SD rules. Consequently, the results revealed by the four evaluation measures indicate a better or similar performance compared to some standard methods.

In addition, the algorithm allows users not only to pick the rules in terms of a rule quality measure, but also according to their associations to topics. The similarity of topics (hence rules) can be visualized by dendograms using the suggested rule distance measure. Last, but not least, CLDA assigns a probability to each feature value in a discovered rule regarding the respective topic, which could aid users in gaining deeper insights into the data.

ACKNOWLEDGEMENTS

The first author acknowledges the support of the TUM Graduate School of Information Science in Health (GSISH), Technische Universität München.

REFERENCES

- [1] D. Andrzejewski, *Incorporating Domain Knowledge in Latent Topic Models*, Ph.D. dissertation, University of Wisconsin-Madison, 2010.
- [2] D. Andrzejewski, X.J. Zhu, and M. Craven, ‘Incorporating domain knowledge into topic modeling via Dirichlet forest priors’, in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 25–32, (2009).
- [3] M. Atzmueller and F. Mitzlaff, ‘Efficient descriptive community mining’, in *24th International Conference of the Florida Artificial Intelligence Research Society*. AAAI Press, (2011).
- [4] D.M. Blei, Ng. Andrew, and M.I. Jordan, ‘Latent Dirichlet allocation’, *Journal of Machine Learning Research*, **3**, 993–1022, (2003).
- [5] M. Boley and H. Grosskreutz, ‘Non-redundant subgroup discovery using a closure system’, in *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 179–194. Springer, (2009).
- [6] V. Dzyuba and M. van Leeuwen, ‘Interactive discovery of interesting subgroup sets’, in *Proceedings of the 12th International Symposium on Intelligent Data Analysis*, (2013).
- [7] K. Farrahi and D. Gatica-Perez, ‘Discovering routines from large-scale human locations using probabilistic topic models’, *ACM Transactions on Intelligent Systems and Technology*, **2**(1), (2011).
- [8] A. Frank and A. Asuncion, ‘UCI machine learning repository’, (2010).
- [9] T.L. Griffiths and M. Steyvers, ‘Finding scientific topics’, *Proceedings of the National Academy of Sciences*, **101**(Suppl. 1), 5228–5235, (2004).
- [10] H. Grosskreutz, S. Rüping, and S. Wrobel, ‘Tight optimistic estimates for fast subgroup discovery’, in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 440–456. Springer, (2008).
- [11] R.F. He and J. Shapiro, ‘Bayesian mixture models for frequent itemset discovery, arxiv, abs/1209.6001’, (2012).
- [12] G. Heinrich, ‘Parameter estimation for text analysis’, Technical report, (2004).
- [13] H.S. Kim, Y.Z. Sun, J. Hockenmaier, and J.W. Han, ‘ETM: Entity topic models for mining documents associated with entities’, in *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, pp. 349–358, (2012).
- [14] W. Klösgen, ‘Explora: A multipattern and multistrategy discovery assistant’, in *Advances in knowledge discovery and data mining*, (1996).
- [15] N. Lavrač, B. Kavsek, P. Flach, L. Todorovski, and S. Wrobel, ‘Subgroup discovery with CN2-SD’, *Journal of Machine Learning Research*, **5**, 153–118, (2004).
- [16] F. Lemmerich, M. Becker, and F. Puppe, ‘Difference-based estimates for generalization-aware subgroup discovery’, in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 288–303. Springer, (2013).
- [17] R. Li, , R. Perneczky, A. Drzezga, and S. Kramer, ‘Efficient redundancy reduced subgroup discovery via quadratic programming’, *Journal of Intelligent Information Systems*, (2013).
- [18] M. Mampaey, J. Vreeken, and N. Tatti, ‘Summarizing data succinctly with the most informative itemsets’, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **6**(4), 16:1–16:42, (2012).
- [19] S. Moens and B. Goethals, ‘Randomly sampling maximal itemsets’, in *IDEA: KDD Workshop on Interactive Data Exploration and Analysis*. ACM, (2013).
- [20] P. K. Novak, N. Lavrač, and G. I. Webb, ‘Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining’, *Journal of Machine Learning Research*, **10**, 377–403, (2009).
- [21] J. Pei, J.W. Han, and W. Wang, ‘Constraint-based sequential pattern mining: the pattern-growth methods’, *Journal of Intelligent Information Systems*, **28**(2), 133–160, (2007).
- [22] M. Scholz, ‘Knowledge-based sampling for subgroup discovery’, in *Local Pattern Detection, Vol. Lecture Notes in Artificial Intelligence 3539*, pp. 171–189. Springer, (2005).
- [23] Y.W. Teh, ‘Dirichlet process’, *Encyclopedia of Machine Learning*, 280–287, (2011).
- [24] M. van Leeuwen and A. Knobbe, ‘Non-redundant subgroup discovery in large and complex data’, in *Proceedings of the 21st European Conference on Machine learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 459–474. Springer, (2011).
- [25] M. van Leeuwen and A. Knobbe, ‘Diverse subgroup set discovery’, *Data Mining and Knowledge Discovery*, **2**(25), 208–242, (2012).
- [26] J. Vreeken, M. van Leeuwen, and A. Siebes, ‘Krimp: mining itemsets that compress’, *Data Mining Knowledge Discovery*, **23**, 169–214, (2011).
- [27] X.R. Wang, *Structured Topic Models: Jointly Modeling Words and Their Accompanying Modalities*, Ph.D. dissertation, University of Massachusetts Amherst, May 2009.
- [28] S. Wrobel, ‘An algorithm for multi-relational discovery of subgroups’, in *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, (1997).
- [29] Z.W. Zhai, B. Liu, H. Xu, and P.F. Jia, ‘Constrained LDA for grouping product features in opinion mining’, in *Proceedings of the 15th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, pp. 448–459, (2011).