Knowledgebase Approximation Using Association Rules Aggregation

Pouya Mehrannia*, Behzad Moshiri, Otman A. Basir

Centre for Pattern Analysis and Machine Intelligence, University of Waterloo, Canada

Abstract

This paper introduces knowledgebase approximation and fusion using association rules aggregation as a means to facilitate accelerated insight induction from high-dimensional and disparate knowledgebases. There are two typical observations that make approximating knowledgebases of interest: (1) it is quite often that insights can be derived based from a partial set of the samples, and not necessarily from all of them; and (2) generally speaking, it is rare that the knowledge of interest is contained in one knowledgebase, but rather distributed among a disparate set of unidentical knowledgebases. As a matter of fact, the insights derivable from knowledgebases tend to be uncertain, even if they were to be derived from a wholistic analysis of the knowledgebase. Thus, optimal knowledgebase approximation may yield the computational efficiency benefit without necessarily compromising insight accuracy. This paper presents a novel method to approximate a set of knowledgebases based on association rule aggregation using the disjunctive pooling rule. We show that this method can reduce insight discovery time while maintaining approximation accuracy within a desirable level.

Keywords: knowledgebase approximation, knowledge Integration, information fusion, association rule mining, disjunctive pooling, Dempster-Shafer theory

1. Introduction

Analyzing data becomes a challenging and expensive task when data starts to grow in volume, variety, and velocity [1]. The accuracy and speed of many of the common predictive techniques degrade on high dimensional data. Abundance of data and high dimensionality may establish a more valuable resource, but entails incorporating more sophisticated predictive analysis [2]. Moreover, the absence of accurate and well-organized data or the incapability of processing large datasets may result in false and spurious insights. Several projection pursuit and manifold methods like principal component analysis (PCA) and multidimensional scaling (MDS) are used for dimensionality reduction for high dimensional data. However, such methods typically rely on assumptions such as the fact that variables are highly correlated or take only numeric values.

Typically, the underlying knowledge of a dataset is more important than the dataset itself in designing information systems [3]. The knowledge extracted from a dataset is stored in knowledgebases which contain information at a higher level of abstraction. Knowledgebases store general facts and rules which might be deduced from thousands of data samples. Therefore, the memory requirements for a knowledgebase is much lower compared to a conventional database.

Creating knowledgebases from datasets of reasonable size is simple, but the complexity of knowledgebase generation grows exponentially with the size of the feature space, especially when typical dimensionality reduction methods are not applicable [4]. In the presence of high dimensional data, it may not be feasible to apply complicated functions directly to the dataset. As a result, many organizations refrain from using some fundamental features to avoid increased computational complexity while those features can enrich induced insights dramatically.

In the majority of cases, all the features that are needed to create a complete and comprehensive knowledgebase cannot be found in a single dataset [5]. While data integration can be used

*Corresponding author

to unify disparate datasets, it does not necessarily construct a reliable dataset containing all the features in one place. Even if it does, the emerging dataset would need a more intensive processing effort to output the desired knowledgebase. As a result, smaller datasets are processed and more and more partial knowledge is produced everyday. Generally speaking, it is rare that the knowledge of interest is contained in one knowledgebase, but rather distributed among a disparate set of unidentical knowledgebases. As such, computational efficiency and knowledge fusion are major design concerns in insight induction systems.

Limitations on allowable processing complexity and available features often require the knowledgebases to be approximated by aggregating the knowledge extracted from smaller datasets. In many cases knowledge approximation results in more accurate insights especially when the knowledge induction process relies on interestingness measures, or when is performed in the presence of noisy and incomplete data. As a matter of fact, the insights derivable from knowledgebases tend to be uncertain, even if they were to be derived from a wholistic analysis of the knowledgebase. Thus, optimal knowledgebase approximation may yield the computational efficiency benefit without necessarily compromising insight accuracy. On the other hand, due to explosion of data, data mining methodologies and information retrieval mechanisms are being revolutionized. Therefore, it is essential to find faster mining approaches and gain deeper insight into recorded data to help making more effective decisions.

Using approximation for reducing computational complexity is widely used for probability models and has been around for a long time. Examples of such approximation techniques can be found in [6] for discrete probability distributions and in [7] for probability models. For knowledgebases, the idea has been developed in the form of knowledge compilation or approximate knowledge fusion [8, 9, 10, 11]. Knowledgebase integration was also a popular topic of investigation for web search engines as the required information of users could be distributed in the databases of various search engines [12].

In the past few years, the necessity of knowledgebase approximation and integration is felt more than ever with the explosion of data in various fields of research. For semantic analysis of social media feeds, for example, integration of local knowledgebases is explored and shows improvement in sentiment analysis and detecting events in social media streams [13]. The idea has also been emerged for learning of knowledgebase representations where the feature types are integrated with a combination of neural representation learning and probabilistic product of experts models [14].

This paper presents a novel approach based on disjunctive rule of combination to approximate knowledgebases. The knowledge here is represented in the form of rules extracted using association rule mining (ARM) [15] techniques. To demonstrate the capacities of knowledgebase approximation, we apply this method to well-known classification problems and show that it successfully generates rules that approximate correlations in the input dataset. This behavior is beneficial for knowledge fusion from multiple datasets and enhances computational efficiency when dealing with high dimensional data.

2. Methodology

The knowledgebase approximation method proposed in this paper uses association rules to represent knowledge. Association rule mining techniques are the tools that extract association rules, and this study incorporates them for knowledge induction. Association rule mining is widely known as a tool for market basket analysis [16], but is applicable to a wide variety of datasets in different domains such as medical diagnosis [17], bioinformatics [18], web mining [19], and traffic accident analysis [20]. This section presents the methodology for our proposed knowledgebase approximation technique on the basis of combination of evidence in the form of association rules.

2.1. Knowledge induction

Association rule mining (ARM) is a fundamental data mining technique with the ability to uncover hidden relationships in a rational dataset while the data might seem unrelated. The set of association rules generated by ARM divulge insights describing the underlying patterns in a dataset. An association rule is denoted by A $\rightarrow$ B with A and B being sets of attributes known respectively as the antecedent and the consequent of the rule. A rule A $\rightarrow$ B is an if/then statement indicating that if A happens then B also happens. A and B are sometimes referred to as itemsets since they can contain a set of disjoint items.

The association of dataset's attributes is mainly found by determining how frequent they appear together in a dataset. Several algorithms have been proposed for this purpose to mine frequent itemsets. The most commonly used algorithms are Apriory, Eclat, and FP-growth [21, 22, 23]. The frequent itemsets are then converted into associations rules by identifying the antecedent and consequent sets.

The three main measures of significance, based on which the ARM applies constrains and selects the interesting rules, are support, confidence, and lift. Equations 1, 2 and 3 respectively calculate the support, confidence, and lift of a rule A $\rightarrow$ B in a dataset. Support of a rule A $\rightarrow$ B indicates the percentage of all itemsets in which A and B occurred together. Confidence of that rule indicates the percentage of itemsets in which when A occurs then B also occurs. Therefore, the former is used to find the frequent itemsets and the latter to filter the strong rules. Lift is also used in the ARM algorithms to measure the correlation between A and B in a dataset. A higher lift value (greater than 1) indicates that A and B appear together more frequently, whereas a lower value (lower than 1) shows reverse of this concept. A lift value 1 would imply that A and B are two independent events and no association rule can involve both events together.

$$Support = P(A \cap B) \tag{1}$$

$$Confidence = P(A|B) = \frac{P(A \cap B)}{P(A)} \tag{2}$$

$$Lift = \frac{P(A \cap B)}{P(A)P(B)} \tag{3}$$

Interesting rules are typically identified based on a minimum support threshold and a minimum confidence threshold. If a rule satisfies both thresholds it is considered as interesting. The thresholds are set empirically in a way that the number of useful rules is maximized and that of the useless ones minimized.

2.2. Combination of evidence

Dempster-Shafer (DS) theory, often described as an extension of the probability theory or a generalization of the Bayesian inference method, offers an alternative for mathematical representation of epistemic uncertainty. As opposed to the traditional probability theory, where evidence is associated with single events, DS theory deals with evidence associated with sets of events and probability values assigned to sets of possibilities. DS theory works at higher levels of abstraction by adding a third aspect, called unknown, to the crisp logic. The basic idea is built upon obtaining degrees of belief from subjective probabilities and combining them using their independent items of evidence [24].

The three main functions used in the DS theory are the basic probability assignment function (BPA or m), the Belief function (Bel), and the Plausibility function (Pl). The BPA function assigns masses to all subsets of the entities in a system by mapping contents of the power set (P_Ω) to the interval between 0 and 1. The mass of subset p_i is commonly denoted by $m(p_i)$ and represents the amount of knowledge associated with that subset. In other words, $m(p_i)$ expresses the proportion of all available evidence that supports p_i but no particular subset of it. Each element $p_i \in P_\Omega$ is called a focal element of P_Ω if $m(p_i) > 0$, and the set of all focal elements is named a body of evidence (BOE). The following three equations represent the above description of m:

$$m : P_\Omega \rightarrow [0, 1] \tag{4}$$

$$m(\varnothing) = 0 \tag{5}$$

$$\sum_{p_i \in P_\Omega} m(p_i) = 1 \tag{6}$$

When multiple independent BOEs are available, which assumes existence of independent generic sources of information, we can use Dempster's rule of combination (DRC) to compute the aggregated BPA on p_i. Having two independent events p_a and p_b with their BPAs expressed by $m_1(p_a)$ and $m_2(p_b)$, DRC can be applied as follows:

$$m_1 \oplus m_2(p_i) = \begin{cases} 0, & \text{for } p_i = \varnothing. \\ \frac{1}{1-k} \sum_{p_a \cap p_b = p_i} m_1(p_a) m_2(p_b), & \text{otherwise.} \end{cases} \tag{7}$$

where

$$k = \sum_{p_a \cap p_b = \varnothing} m_1(p_a) m_2(p_b)$$

is a normalization constant called conflict degree and represents the amount of conflicting evidence between the two sources of information. DRC is purely a conjunctive operation which is AND-based

and operates on set intersection. In the situation where not every source is reliable and at least one reliable source exists, a modified DRC, known as disjunctive pooling rule (DPR) [25], is more appropriate. As opposed to DRC, DPR is OR-based and operates on set union. DPR does not reject any of the information asserted by the sources and does not generate any conflict. It can be applied to two independent events p_a and p_b using Equation 8:

$$(m_1 \boxplus m_2)(p_i) = \sum_{p_a \cup p_b = p_i} m_1(p_a) m_2(p_b) \qquad (8)$$

The other two key functions in the DS theory, the Belief and Plausibility functions, are two non-additive continuous measures perceived as the lower and upper bounds of the interval containing the exact probability at which p_i is supported [26]. Both functions are calculated based on basic probability assignment as indicated in Equations 10 and 11. The lower bound, Belief, is defined as the sum of all the masses of subsets of the set of interest, whereas the upper bound, Plausibility, is the sum of all the masses of the sets that intersect the set of interest.

$$Bel(p_i) \leq P(p_i) \leq Pl(p_i) \qquad (9)$$

$$Bel(p_i) = \sum_{p_k | p_k \subseteq p_i} m(p_k) \qquad (10)$$

$$Pl(p_i) = \sum_{p_k | p_k \cap p_i = \varnothing} m(p_k) \qquad (11)$$

The two measures of Belief and Plausibility can be derived from each other by the following relations:

$$pl(p_i) = [1 - Bel(\overline{p_i})] \qquad (12)$$

$$Bel(p_i) = [1 - Pl(\overline{p_i})] \qquad (13)$$

where $\overline{p_i}$ denotes the complement of p_i.

DPR is more robust than DRC in the presence of conflicting evidence, and its use is appropriate when the conflict is due to poor reliability of some of the sources. In other words, DRC works based on the assumption that the belief functions to be combined are induced by reliable sources of information, whereas the DPR only assumes that at least one source of information is reliable, but we do not know which one. Both rules assume the sources of information to be independent. DPR is defined based on the union of the basic probability assignments (BPA) by extending the set-theory union and hence is an appropriate operator for insight aggregation. Some other characteristics of DPR that recommend it for this purpose are:

- Unlike conjunctive pooling, disjunctive pooling incorporates all the information asserted by the sources rather than selecting the part which is in consensus.
- The union does not generate any conflict
- No normalization procedure is required
- DPR is commutative and associative, but not idempotent
- The belief measure associated with aggregated BPAs is easily calculated via multiplication of individual BPA belief measures, i.e., $Bel(p_i) = Bel_1(p_i)Bel_2(p_i)$

2.3. Knowledgebase approximation framework

The focus of our approach for knowledgebase approximation is on integration of knowledge, which is drawn in the form of if/then rules using the ARM method, from smaller datasets with fewer features. These smaller datasets may be obtained from different data providers with their own objectives. In this case, approximating the knowledgebase corresponding to the integrated dataset would save the hassle of dataset integration and processing the bigger emerging dataset. Nonetheless, any large dataset can be broken into smaller ones by selecting only certain features to appear in each of them. As a result, we just need to deal with multiple lower-dimensional datasets requiring lower computation efforts.

Let us assume that N independent datasets are available for investigation indicated by DB_i and $i \in \{1, 2, \cdots, N\}$. These datasets are the main sources of information from which we aspire to obtain an approximated knowledgebase comprising all the attributes appeared in any of the datasets. Any pair of the datasets may share common features. Hence, the dimension of the corresponding integrated dataset is not necessarily equal to the sum of the number of features. At the end of this section we will show that common features help the DPR method to find the best approximation.

In order to induce knowledge from the smaller datasets, ARM is applied to each dataset which

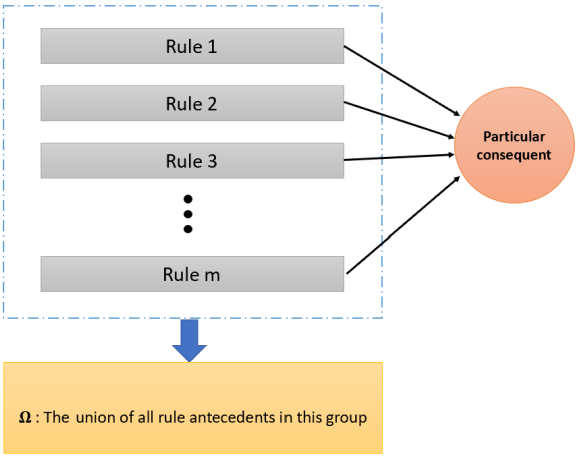


Figure 1: Rules with the same consequent are filtered to be used for approximating a bigger piece of knowledge

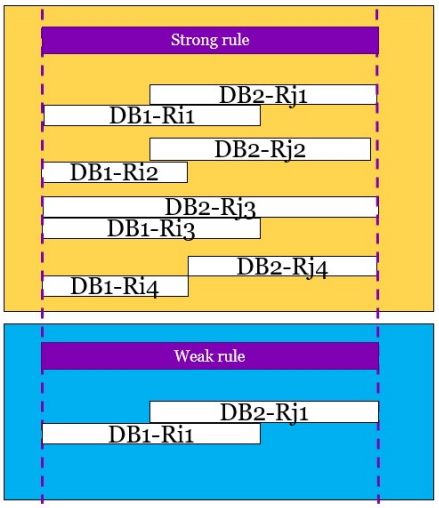


Figure 2: Strong vs. weak rules in DPR

generates N independent rulesets. ARM explores and connects the attributes that contribute in the occurrence of a particular event or a set of events. Depending on the size and nature of the datasets, different ARM methods can be used. Given a minimum support threshold and a minimum confidence threshold, ARM finds all the strong association rules, that is, those whose confidence and support values are equal or greater than the thresholds. A rule that does not meet the thresholds is called a weak association rule.

Having mined all the association rules from the available datasets, there will be N independent rulesets available which are denoted by RS_i and $i \in \{1, 2, \cdots, N\}$. Each rule in any of the rulesets can be regarded as a piece of evidence in the way defined in the DS theory and can be assigned a mass value. This mass value can be simply a weighted average of the ARM interestingness measures including support, confidence, and lift. Since the summation of mass values for a specific event should equal to 1, these weighted averages are normalized over all the rules that point to a particular consequent. More sophisticated BPA mappings can be defined depending on the degree to which domain knowledge is available.

Since ARM restricts the rules to those satisfying minimum support and confidence thresholds, masses corresponding to the emerging rules can be assumed to have non-zero BPAs and hence regarded as focal elements. Consequently, by mapping the interestingness of the rules to mass values, the rulesets can be transformed into independent bodies of evidence (BOEs) to which combination rules can be applied. In other words, the mass value assigned to a focal element is proportional to its generating rule's strength or interestingness.

This paper incorporates DPR to combine the independent BOEs obtained from the lower dimensional datasets. DPR is a union-based operator and unlike DRC, which selects a condensed part of evidence, DPR selects an extended piece of evidence based on the number and weights of the BOEs that can shape that extended piece of evidence in aggregate. In our case, pieces of evidence represent association rules and extending them will generate a rule with a larger number of antecedents. To merge the antecedents of multiple rules, their consequents should be the same, as shown in Figure 1 . Therefore, the rules are filtered into groups in advance based on their consequents and the process of rule to BOE transformation and applying DPR is performed for each group separately.

As illustrated in Equation 8, DPR uses the values of BPAs in different domains to find a fused set of masses assigned to the higher dimensional domain. The equation implies that the BPA for a fused rule is calculated by summing over the multiplication of those masses whose rule's antecedents can merge into the antecedents of that fused rule. For example, if {A and B $\rightarrow$ P} and {B and C $\rightarrow$ P} then masses for these two rules will be used to calculate the BPA for the fused rule {A and B and C $\rightarrow$ P}.

In a simplified version of the problem when the

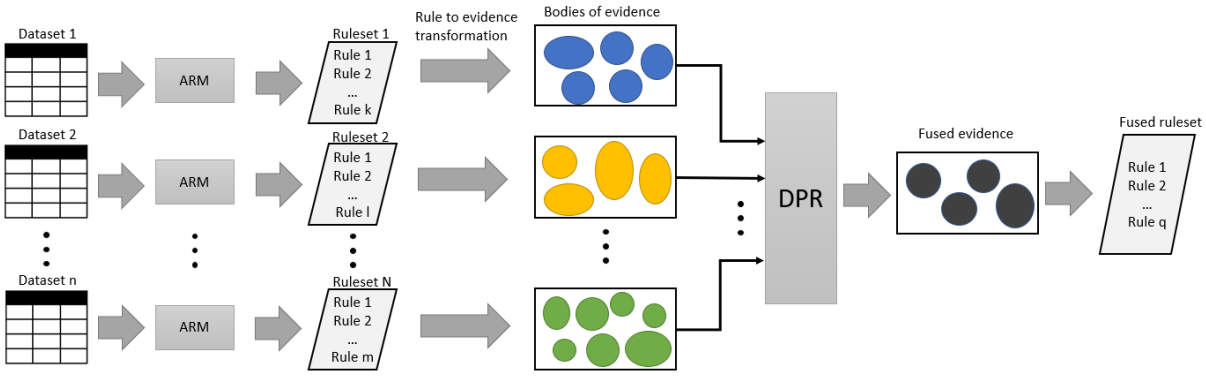


Figure 3: Application of DPR for knowledgebase approximation

BPA values are disregarded, the strength of association rules are dismissed and all rules will have the same impact on generation of the elements in the fused set. Figure 2 elaborates on how DPR differentiates between strong and weak extended rules when BPA values are not considered. In this figure, $DB_q - R_k$ represents rule number k induced from DB_q and $DB_1 - R_{i1}$ is the first rule found in DB_1 that can generate a specified extended rule in aggregate with another rule $DB_2 - R_{i2}$ found in DB_2. In this figure, strong rule is an extended rule which is reproduced a good many times by the union of different pairs of rules, one from DB_1 and one from DB_2. In contrast, weak rule is reproduced relatively infrequent, which means not many rules from DB_1 and DB_2 could be augmented in their antecedents to form that extended association rule.

In the real scenario, the BPA values are not disregarded and each association rule in the rulesets is assigned a mass corresponding to the rule's strength. What changes in this case is that the strength of extended association rules is not measured only based on the count of reproduction times. Instead, the multiplications of the masses for any pairs of rules whose antecedents' union can reproduce the specified extended rule are accumulated. This procedure can also be applied to more than two datasets where the union should be conducted on a combination of rules from disparate datasets. Figure 3 outlines the proposed method in which N independent datasets are assumed available for investigation.

Mining transaction datasets for association rules typically generates a large number of rules. When ARM is used for subsequent prediction, most of the rules become unnecessary and can be eliminated. In our method, however, we utilize every generated rule that satisfies the minimum support and confidence thresholds. Using the informative rulesets along with their dependant rules helps the DPR method to better sort the extended association rules based on their strength. When the fused rules and their corresponding BPAs are created then we can keep the informative ruleset and eliminate the dependant rules.

Rules' masses in our method are obtained by multiplication of their support and confidence, and normalizing them over the whole ruleset. Let us assume q rules are mined from a ruleset. If the rule r_i has the confidence $conf(r_i)$ and support $sup(r_i)$, then its mass is calculated using Equation 14.

$$m(r_i) = \frac{conf(r_i) \times sup(r_i))}{\sum_{j=1}^{q} conf(r_j) \times sup(r_j))} \tag{14}$$

When all the rules in N rulesets transformed into BOEs, DPR can be applied to integrate them into a single set of probability mass assignment indicated by fused evidence in Figure 3. This set is a combination of masses attributed to both strong and weak rules, but we can easily prune the masses and keep the stronger ones. To do so, we should consider the BPA in the fused set as a measure that is proportional to the multiplication of confidence and support of a rule that generated it. We refer to this measure as rule strength. Those BPAs in the fused set that can generate rules with a strength more than a minimum threshold will be selected as dominant BPAs and will be used to generate the integrated insights. The minimum strength thresh-

Table 1: Lyphography dataset features

Attribute	Attribute description and values
Lymphatic	A test for the overall lymphatic system: value 1 for normal; Value 2 for arched, value 3 for deformed; and value 4 for displaced
Block of afferent	value 1 for no; value 2 for yes;
Block of lymph c	value 1 for no; value 2 for yes;
Block of lymph s	value 1 for no; value 2 for yes;
By pass	value 1 for no; value 2 for yes;
Extravasates	expel from a vessel and is represented by 1 and 2;
Regeneration	value 1 for no; value 2 for yes;
Early uptake	value 1 for no; value 2 for yes;
Lymph nodes dimension	ranges from 0 to 3;
Lymph nodes enlarge	ranges from 1 to 4;
Changes in lymph	value 1 for bean, value 2 for oval and value 3 for round;
Defect in node	value 1 for no, value 2 for lacunars, value 3 for lacunars marginal and value 4 for lacunars central
Changes in node	value 1 for no, value 2 for lacunars, value 3 for lacunars marginal and value 4 for lacunars central;
Changes in structure	the structure of the lymphatic system; values 1 to 8, respectively, for: no, grainy, drop-like, coarse, diluted, reticular, stripped, and faint
Special forms	value 1 for no, value 2 for chalices and value 3 for vesicles;
Dislocation of node	value 1 for no and value 2 for yes;
Exclusion of node	value 1 for no and value 2 for yes;
Number of nodes	Values 1 to 7 for the number of nodes in the range of 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, and 60-69; and value 8 for equal or greater than 70

old (MST) for this purpose is calculated using the minimum confidence and support of the rules in the original rulesets as indicated in Equation 15.

$$MST = Min(conf) \times Min(sup) \qquad (15)$$

3. Application to pattern recognition

Associative classification is an integration of association rule mining and classification which has been investigated widely in pattern recognition. Previous studies show that associative classification can achieve a high classification accuracy and is highly flexible at handling unstructured data. Among the algorithms proposed for classification based on multiple-class association rules CMAR and CPAR are shown to have competitive performance based on the experimental results in [27], [28]. In this paper we apply CMAR to the association ruleset in an approximated knowledgebase that is obtained by our association rule aggregation method and show that the accuracy of classification can be maintained when certain number of attributes are in common between two datasets. In other words, a knowledgebase can be approximated by applying our method to its corresponding lower dimensional datasets when there is enough information shared among them.

We have used the lymphography dataset [29], along with 6 other datasets obtained from UCI (university of California Irvine) machine learning database, to evaluate our DPR-based knowledge approximation framework. We use the lymphography dataset as an example to show how DPR based approximation is applied to a dataset, but we will use all 7 datasets to show effectiveness of our approach in terms of approximation accuracy and run time.

The lymphography dataset was recorded at the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. It contains 148 instances and 19 numerical valued attributes related to different aspects of lymphographic clinical data including the class attribute. Table 1 describes these attributes and the values used for them in the dataset. The class variable holds any of the four cases of normal, metastases, malign lymph, and fibrosis.

As described in the previous section, this paper adopts ARM for extracting knowledge in the form of association rules. One of the criteria in ARM to select interesting rules is support which targets those rules that their components appear in the dataset adequately. Among the existing classes in the lymphography dataset two classes of normal and fibrosis contain very few samples and cannot satisfy the support measure. Therefore, our investigation is limited to the classes of metastases and malignant.

In order to approximate the knowledgebase for this dataset, we divided it into two smaller datasets

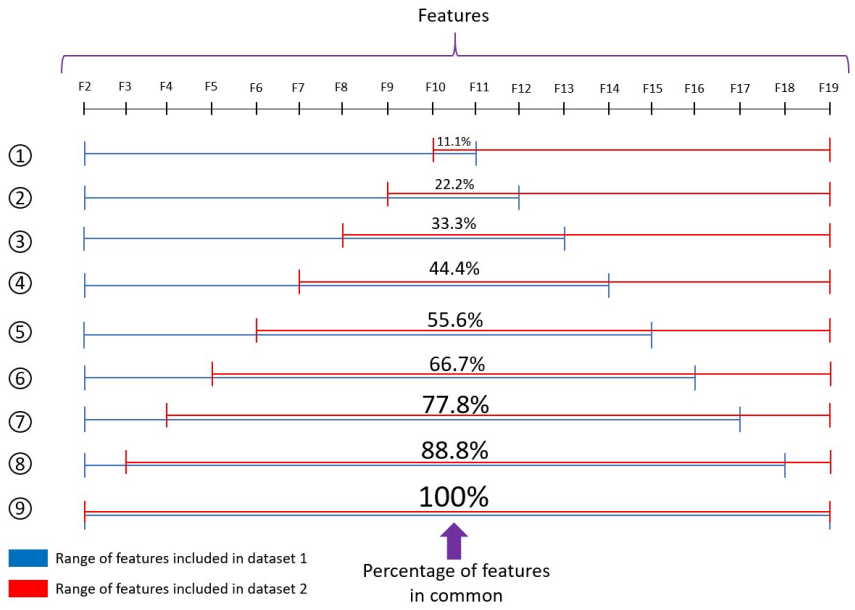


Figure 4: Exploring the effect of number of in-common features in the Lymphography dataset

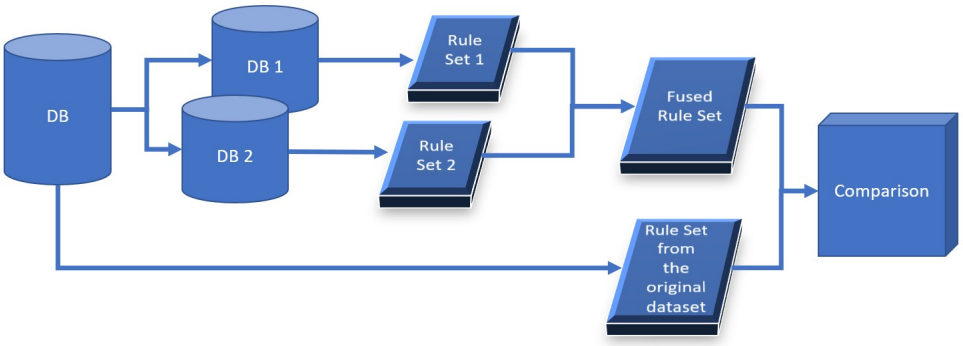


Figure 5: Evaluation framework

in the feature space. This division was done 9 times and each time the percentage of common features was increased. Figure 4 shows this procedure where the percentage of common features is increased by 11.1 (2 features out of 18) in the range of 11% to 100%. The original dataset contains 18 features indicated by F2 to F19. As an example, when the percentage of features in common is 55.6%, the smaller datasets hold 14 features each where F2 to F15 are in one and F6 to F19 are in the other. F6 to F15 are the features in common between the datasets in this example.

Once the original dataset is divided, ARM is applied to each smaller dataset to extract rulesets as illustrated in Figure 5. Before proceeding with integration, the rules were filtrated to separate the rules with consequences of metastases and malignant as shown in Figure 1. The smaller rulesets with matched consequents are then independently aggregated base on our proposed DPR-based integration framework to create two sets of extended rules with metastases and malignant as the consequents. These two sets can then be blended as a single set indicated by fused rules to represent the information in the approximated knowledgebase.

As shown in Figure 5, the ruleset from the original dataset was also induced as the ground truth for evaluating the proposed approximation method in this case study. As mentioned at the beginning of this section, the accuracy of associative classifi-

Table 2: Approximation accuracy

Percentage of in-common features	CMAR classification accuracy	Approximation accuracy	processing time (ms)
11.1%	28.3%	34%	6
22.2%	41.8%	50.3%	12
33.3%	52%	62.5%	17
44.4%	61.2%	73.6%	26
55.6%	63.9%	76.9%	33
66.7%	69%	83%	48
77.8%	74.2%	89.2%	89
88.8%	79.3 %	95.4%	177
100%	83.1%	100%	1073

cation is used to compare the knowledgebase and its approximation. CMAR classification method is applied to the fused ruleset and the ruleset induced directly from the original ruleset to compare the classification accuracy. For the classification purpose, 10-fold cross validation was used with the 70% percent of the original dataset's samples for training.

The same approach is applied to the other 6 UCI datasets and reported in the Section 4. The DPR-based knowledgebase approximation can also be adopted to more than 2 datasets. In this case, there are two approaches available that have the same outcome. One is to apply the fusion operation to all datasets at the same time, an the other is to integrate rulesets in a cascade manner, i.e., the resultant approximated ruleset of each two is fused with the next ruleset.

4. Results and discussion

In this study, the Apriori algorithm is used, which is the best-known ARM method and a simple approach for extracting association rules in a cohort dataset. However, other algorithms could be incorporated to generate insights for the proposed insight aggregation method. One important specification of the Apriory algorithm is that it uses a bottom up approach, i.e., one item is added to the frequent itemsets at a time and tested against the data. The breadth-first search nature of this algorithm makes it suitable for finding desired rules without considerable computation complexity from the small dataset in use here. Furthermore, it is easier to store the dependant rules of those in the smallest informative ruleset.

Based on empirical experiments, we set the value of support to 25%. A reasonable value for support is essential for identifying the rules worth considering for further analysis. If an itemset happens to have a very low support, it will not provide enough information on the relationship between its items, and hence no concrete conclusion can be drawn from such a rule.

The number of instances in the lymph dataset is 148 and a support of 25% guarantees that selected itemsets show up together in at least 37 instances. Although we can find many rules with a confidence value of 100% if we decrease the support threshold, those rules are not necessary useful since there are not enough occurrences of their itemsets. To better understand this fact, suppose there are only 2 instances of items A, B, and C happening together. It is not infrequent that a rule with confidence of 100% is derived from it. In contrast, if the number of instances is high (like 10K or more) then it makes sense to lower the support because even 1% in such case still has many (1000 and more in the 10k example) instances.

As the goal of lymph classification was to relate predictor variables to the occurrence of two classes of metastasis and malignant, all predictor variables are limited to appear only in the antecedent (IF part), and lymph classes (outcome variable) to appear only in the consequent (THEN part). To generate all the strong association rules, we conducted our analysis by selecting any rule satisfying initial support threshold of 25% and confidence threshold of 40% for the generation of frequent itemsets and rule induction. For each of the two smaller datasets, two types of rules were extracted for metastasis and malignant, separately. We used orange3 in python

Table 3: Metastasis

Percentage of in-common features	Number of key features in common	Number of correct recovered rules	percentage of correct recovered rules	Number of incorrect recovered rules	percentage of incorrect recovered rules
11.1%	2	11	36%	82	273%
22.2%	4	15	50%	65	217%
33.3%	5	16	54%	36	120%
44.4%	7	20	67%	22	74%
55.6%	8	23	77%	13	44%
66.7%	9	25	84%	8	27%
77.8%	11	26	87%	6	20%
88.8%	13	29	96%	2	6%
100%	14	30	100%	0	0%

Table 4: Malignant

Percentage of in-common features	Number of key features in common	Number of correct recovered rules	percentage of correct recovered rules	Number of incorrect recovered rules	percentage of incorrect recovered rules
11.1%	1	4	21%	16	84%
22.2%	3	6	32%	14	74%
33.3%	4	8	42%	11	58%
44.4%	4	8	42%	11	58%
55.6%	5	11	58%	7	37%
66.7%	7	15	79%	3	16%
77.8%	9	19	100%	1	5%
88.8%	11	19	100%	0	0%
100%	11	19	100%	0	0%

3.4 to apply the association rule algorithm.

After approximating the original dataset's knowledgebase by integrating the association rules from the smaller datasets, CMAR associative classification was applied to the aggregated rules. We applied CMAR to the variation of in common features and used 10-fold cross validation as the original data samples were limited . The accuracies are reported in the table 2. The accuracy of CMAR in classifying the original dataset is 83.1% which is used to normalize the approximation accuracy. In other words, an approximation accuracy of 100% indicates that the maximum possible classification accuracy is obtained, i.e. 83.1%.

As discussed in the previous sections, the processing time for finding association rules, or in general inducing insights, increases exponentially by the size of feature space. Our approximation method decreases this processing time significantly while the approximated knowledgebase is highly accurate when there are enough features in common. The average processing time for the nine experiments reported in table 2 using our method is 0.16s while the run time of extracting association rules from the original dataset is about 1.1s. It is worthwhile to mention that the lymph dataset is a relatively small dataset containing only 143 samples, 18 features and 4 classes. The run time in our method is ten times less. It is trivial that using our method for bigger datasets can save much more time.

Tables 3 and 4 report the number of correct and incorrect recovered rules in each experiment for the consequents of metastasis and malignant separately. The basis to distinguish correct and incorrect rules is the smallest informative ruleset induced from the original dataset. This basis is selected because our initial goal was to approximate the knowledgebase from the original dataset and choosing this basis conforms to our goal. In these tables, key features are those with higher relevancy to the consequents and can better distinguish the outcome. In the lymph dataset, 7 features exist that can be used as key features to differentiate between a metastasis and a malignant lymph.

Table 5 shows the summery of results for all 7 UCI datasets when our approximation method is applied. In this table, acc denotes accuracy and RT denotes run time. The run time and accuracy

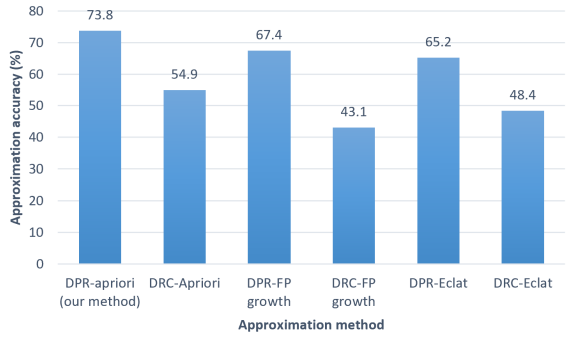


Figure 6: Comparison of approximation accuracy for 6 different approximation methods

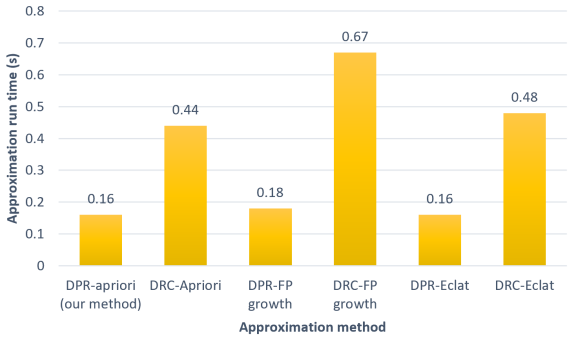


Figure 7: Comparison of approximation run time for 6 different approximation methods

of the approximation method is the average of run times for all 9 experiments in which the number of common features are increased from 10% to 100% in step sizes of 10%. As can be seen from the table, the averaged approximation accuracy drops when the CMAR accuracy increases. The reason is that the CMAR accuracy is used to normalize the approximation accuracy, and hence, the real accuracy of classification on the approximated dataset remains in the acceptable range of 70 to 90 percent. Another considerable observation is the run time of the classification on the approximated knowledgebase which is significantly lower than the CMAR run time on the original dataset.

We compared the approximation accuracy and run-time of our proposed method with those of 5 other approximation methods. These 5 methods follow the same procedures described in this paper, but differ from our proposed method in either using DCR instead of DPR to combine the BPAs, or using a different ARM method to generate the frequent itemsets. Figures 6 and 7 illustrate the accuracies and run-time of these methods. As explained in Section 2.2, DRC is a conjunctive operation which is AND-based and operates on set intersection. DRC works based on the consensus of the evidence and hence reports the the antecedents of the fused rules as the items that are in agreement with many of the rules in the initial datasets. This makes the fused rules to shrink, which is not ideal for integration of knowledge, and lowers the accuracy of knowledge approximation which can be seen in Figure 6. It also affects negatively on the run-time as it has to compute rule conflicts every time a fused BPA is being generated.

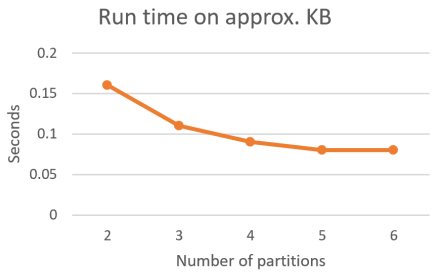
Some of the other methods in Figures 6 and 7 use Frequent Pattern (FP) Growth or Equivalence Class Clustering and bottom-up Lattice Traversal (Eclat) to generate their association rules. FP-growth algorithm is the method of finding frequent patterns without candidate generation. It constructs an FP Tree rather than using the generate and test strategy of Apriori. Eclat, on the other hand, uses Transaction Id Sets(tidsets) intersections to compute the support value of a candidate and prevents the generation of subsets which do not exist in the prefix tree. According to figure 6, our DPR-apriori method outperforms both FP-growth and Eclat methods no matter which combination rule they are being used with. The Eclat method, however, competes with the apriori method in the run-time (Figure 7) since it is an efficient ARM method which works in a vertical manner similar to a Depth-First Search strategy in a graph.

As previously noted in Section 3, the DPR-based knowledgebase approximation can be applied to more than 2 datasets and there are two approaches to apply the fusion operation with the same outcome. Breaking the dataset into more partitions can boost the processing speed even further by enabling parallel computation on these partitions. We investigated the effect of number of partitions on the run time and accuracy of classification using the approximated Lymph dataset. First, we increased the number of partitions when all partitions are in use and monitored the run time and accuracy. Second, we fixed the number of partitions to 6 and started using only a specific number of those 6 partitions in the range of 2 to 6. The results of these two cases are shown in Figure 8 and Figure 9 respectively.

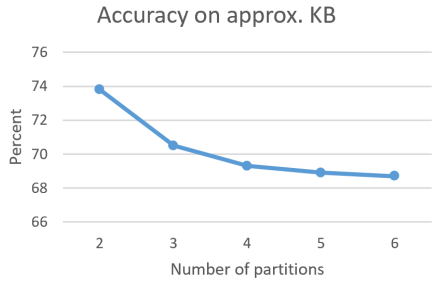
In the first case, the approximation accuracy starts to drop as the number of partitions increases. The reason is that the approximation takes place

Table 5: The comparison of CMAR accuracy and run time when applied to the original dataset and the approximated knowledgebase

dataset	# attr	# cls	# rec	CMAR acc	CMAR RT	approx acc	approx RT
Lymph	18	2	148	83.1%	1.1s	73.8%	0.16s
Auto	25	7	205	78.1%	4.4s	86.1%	0.34s
Hypo	25	2	3163	98.4%	7.9s	69.8%	0.48s
Iono	34	2	351	91.5%	3.5s	75.8%	0.26s
Sick	29	2	2800	97.5%	11.7s	72%	0.5s
Sonar	60	2	208	79.4%	11.1s	81.3%	0.52s
Vehicle	18	4	846	68.8%	1.6s	85%	0.18s

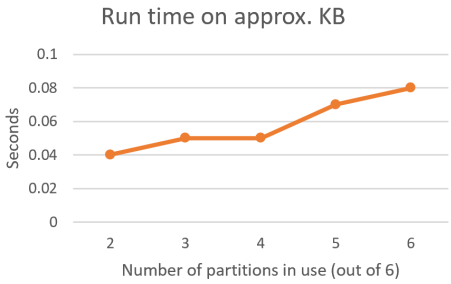


(a) effect on run time

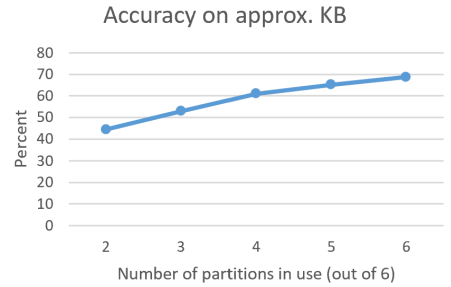


(b) effect on accuracy

Figure 8: Experiment on the Lymph dataset to show the effect of number of partitions when all partitions are in use



(a) effect on run time



(b) effect on accuracy

Figure 9: Experiment on the Lymph dataset to show the effect of number of partitions in use when total number of partitions is 6

based on smaller pieces of the dataset and hence less information is contained in the rulesets. As a result, The in-common rules become smaller and smaller until only the key features start to drive the classification. This can be seen in figure 8-b when the slope of the connecting lines start to decrease. The same thing happens to the run time. The run time starts to get smaller until there are no more available processing units to extract the rules of new partitions in parallel with the previous partitions.

In contrast, the approximation accuracy and run time increase in the second case as the number of partitions in use increases. This is trivial since more information becomes available when more partitions are in use. The effect of number of partitions in this case are illustrated in Figures 9-a and 9-b. It is worthwhile to mention that the dataset in this case is partitioned against the second dimension, i.e. the samples, as opposed to the previous case where the dataset was partitioned against the feature space.

5. Conclusion

Knowledge discovery is becoming a central issue for industrial and government organizations. The ability of these organizations to conduct their business, effectively and efficiently, is heavily dependent

on insights they derive from knowledgebases relevant to their businesses. This has led to the emergence of insights deduction systems as an important computing discipline. It is typical that Big Data Knowledgebases tend to be disparate with high dimensionality. In the presence of high dimensional data, it may not be feasible to apply complicated functions directly to the dataset. As such, computational efficiency and knowledge fusion are major design concerns in insight induction systems.

This paper introduced a knowledgebase approximation methodology to address two challenges in data analysis: (1) association rule mining efficiency at handling huge datasets, and (2) integration of induced rules from disparate datasets without the need for integration in data level. We proposed using the DPR approach along with the BPA assignment in the Dempster-Shafer theory to combine the rulesets and assign a new measure of interestingness, i.e. rule strength, to the fused rules. The fused rulset is an approximate knowledgebase for the whole data available in disparate datasets. Our experiments on the lymphography and 6 other datasets in the UCI machine learning database repository show that DPR-base approximation can achieve high accuracy when the number of in-common features between two smaller datasets are above 60%. DPR investigates the integration of knowledge as opposed to the consensus of knowledge in DRC and therefore it does not need to manage the conflicts in the rules. However, the impact of conflict in the approximation accuracy can be investigated as a future work.

6. conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] S. Kaisler, F. Armour, J. A. Espinosa, W. Money, Big data: Issues and challenges moving forward, in: 2013 46th Hawaii International Conference on System Sciences, IEEE, pp. 995–1004.

[2] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nature reviews cancer 8 (2008) 37–49.

[3] A. Bordes, J. Weston, R. Collobert, Y. Bengio, Learning structured embeddings of knowledge bases, in: Conference on artificial intelligence, CONF.

[4] L. Van Der Maaten, E. Postma, J. Van den Herik, Dimensionality reduction: a comparative, J Mach Learn Res 10 (2009) 13.

[5] O. Miotto, T. W. Tan, V. Brusic, Rule-based knowledge aggregation for large-scale protein sequence analysis of influenza a viruses, in: BMC bioinformatics, volume 9, Springer, p. S7.

[6] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, IEEE transactions on Information Theory 14 (1968) 462–467.

[7] H. J. Kappen, W. Wiegerinck, Second order approximations for probability models, in: Advances in Neural Information Processing Systems (2001), pp. 238–244.

[8] B. Selman, H. A. Kautz, Knowledge compilation using horn approximations., in: AAAI (1991), Citeseer, pp. 904–909.

[9] P. Z. D. Martires, A. Dries, L. De Raedt, Knowledge compilation with continuous random variables and its application in hybrid probabilistic logic programming, arXiv preprint arXiv:1807.00614 (2018).

[10] B. Dunin-Ke, L. A. Nguyen, A. Szałas, et al., Tractable approximate knowledge fusion using the horn fragment of serial propositional dynamic logic, International Journal of Approximate Reasoning 51 (2010) 346–362.

[11] N. Dangdang, L. Lei, L. Shuai, Knowledge compilation methods based on the clausal relevance and extension rule, Chinese Journal of Electronics 27 (2018) 1037–1042.

[12] A. H. Keyhanipour, B. Moshiri, M. Kazemian, M. Piroozmand, C. Lucas, Aggregation of web search engines based on users' preferences in webfusion, Knowledge-Based Systems 20 (2007) 321–328.

[13] T. Kolajo, O. Daramola, A. Adebiyi, A. Seth, A framework for pre-processing of social media feeds based on integrated local knowledge base, Information Processing & Management 57 (2020) 102348.

[14] A. Garcia-Duran, M. Niepert, Kblrn: End-to-end learning of knowledge base representations with latent, relational, and numerical features, arXiv preprint arXiv:1709.04676 (2017).

[15] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Acm sigmod record (1993), volume 22, ACM, pp. 207–216.

[16] S. Brin, R. Motwani, J. D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, in: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pp. 255–264.

[17] J. Nahar, T. Imam, K. S. Tickle, Y.-P. P. Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Expert Systems with Applications 40 (2013) 1086–1093.

[18] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, O. Gandrillon, Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data, Genome Biology 3 (2002) 1–16.

[19] C.-H. Lee, Y.-H. Kim, P.-K. Rhee, Web personalization expert with combining collaborative filtering and association rule mining technique, Expert Systems with Applications 21 (2001) 131–137.

[20] Mehrannia, Pouya, Temporospatial Context-Aware Vehicular Crash Risk Prediction, Ph.D. thesis, University of Waterloo, 2020.

[21] J. Han, J. Pei, Y. Yin, Mining frequent patterns with-

out candidate generation, in: ACM sigmod record (2000), volume 29, ACM, pp. 1–12.
[22] M. J. Zaki, Scalable algorithms for association mining, IEEE transactions on knowledge and data engineering 12 (2000) 372–390.
[23] R. Agarwal, R. Srikant, et al., Fast algorithms for mining association rules, in: Proc. of the 20th VLDB Conference (1994), pp. 487–499.
[24] A. P. Dempster, A generalization of bayesian inference, in: Classic works of the dempster-shafer theory of belief functions, Springer, 2008, pp. 73–104.
[25] D. Dubois, H. Prade, Representation and combination of uncertainty with belief functions and possibility measures, Computational intelligence 4 (1988) 244–264.
[26] X. Gros, NDT data fusion, Elsevier, 1996.
[27] W. Li, J. Han, J. Pei, Cmar: Accurate and efficient classification based on multiple class-association rules, in: icdm (2001), IEEE, p. 369.
[28] X. Yin, J. Han, Cpar: Classification based on predictive association rules, in: Proceedings of the 2003 SIAM International Conference on Data Mining (2003), SIAM, pp. 331–335.
[29] *UCI dataset*, 1988. `https://archive.ics.uci.edu/ml/datasets` [Accessed: 2020-03-30].