Towards Association Rules as a Predictive Tool for Geospatial Areas Evolution

Asma Gharbi[1,2], Cyril De Runz[3,4,1], Sami Faiz[2] and Herman Akdag[1]
[1]*LIASD, University of Paris 8, Saint-Denis, France*
[2]*LTSIRS, University of La Mannouba, Tunis, Tunisia*
[3]*CReSTIC University of Champagne-Adrenne, Reims, France*
[4]*Sorbonne Université, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, Paris, France*

Keywords: Association Rules, Spatial Dynamics, Prediction, Sequential Association Rules, Class Association Rules.

Abstract: Although it was basically presented as an exploratory tool rather than a predictive tool, numerous follow up researches have enhanced association rule mining, which contributes in making it a powerful predictive tool. In this context, this paper review the main advances in this datamining technique, then attempts to describe how they can, practically, be harnessed to deal with problems such as the prediction of geographical areas evolution.

1 INTRODUCTION

Geographical areas, such as cities are large, complex, and dynamic systems evolving and changing their characteristics under the effect of divers social, economic, political, and environmental factors. The need for understanding, projecting, and planning their evolution, mainly consisting in land use/cover change (LUCC), has emerged decades ago as an attempt to accommodate urban dynamics while preserving and restoring the environment. For instance, in (Gharbi et al., 2014), urban areas have been perceived as collections of geographical entities (GEs) temporally and spatially related to one another and continuously evolving on two main levels: the spatial level related to their morphologies and locations, and the functional level regarding their vocations or use. Indeed, GEs' evolutions have been represented in form of sequences embedding their land use and spatial configuration histories.

In the present paper, we propose a solution based on a widely adopted prediction hypothesis according to which, the best predictor of future is the past. In this context we quote the words of great minds in different domains such as, Robert Kiyosaki, the famous American business man, and financial literacy activist and commentator, who pointed out that: *"The best way to predict the future is to study the past, or prognosticate."*; Jules Henri Poincaré, a French mathematician, theoretical physicist, engineer, and philosopher of science, who said that *"If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of the same universe at a succeeding moment."*; and last, but not least, we may as well quote the clinical psychologist Albert Ellis who asserted that, in the psychology domain, *"The best predictor of future behavior is past behavior"*. The hypothesis above, which constitutes the core of our work, corresponds to a popular datamining approach known as frequent pattern mining (FPM). This approach consists in identifying items, sequences, and the frequently co-occurring structures in a given past database, with the objective of discovering relationships among the different variables describing the data. Such correlations or associations are generally represented in form of rules. Since decades, FPM has known abundant ameliorations and improvements to adequately respond to tasks such as prediction.

In this context, this paper suggests the use of association rule mining (ARM) to extract evolution rules for a certain geographical area, based on time series spatio-temporal data. Therefore, we starts in section 2 by defining the association rule mining problem, presenting the progress it has known over years, how this progress has contributed in making association rules a powerful predictive tool which motivates its employment for spatial dynamic prediction. Thereafter, we dedicate the third section for describing how we used adapted association rules for urban dynamic predic-

Gharbi, A., Runz, C., Faiz, S. and Akdag, H.
Towards Association Rules as a Predictive Tool for Geospatial Areas Evolution.
In *Proceedings of the 2nd International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2016)*, pages 201-206
ISBN: 978-989-758-188-5

tion. Finally, in section 4, we draw a conclusion.

2 ASSOCIATION RULE: DEFINITION AND EVOLUTION

2.1 Problem Statement

Association rule mining is a fundamental datamining task which aims at discovering useful regularities, called associations, in a given database, through identifying all the frequently co-occurring items. The formal definition of association rules can be stated as follows: Let D be a transaction database and T a set of transactions (T=$\{t_1, t_2, t_3, ..., t_n\}$) of D, composed of a set of items I=$\{i_1, i_2, i_3, ..., i_x\}$, such that $t_i \subseteq I$. An association rule represents an implication in the following form:$X \longrightarrow Y$, where X and Y are sets of items, called itemsets; X, $Y \subset I$; and $X \cap Y = \emptyset$.

Since it was first introduced in (Agrawal et al., 1993), association rule mining has been a focused research area in machine learning and knowledge discovery, over decades. Apriori is one of the most popular first presented algorithms. It proceeds in two steps:

1. Generation of all frequent itemsets using a level-wise complete search algorithm. It generates candidate itemsets, based on the downward closure property. Then it scans the database to determine their support values and keep only the frequent ones. An itemset is called frequent if its support count (the number of transactions containing this itemset) is equal or greater than a user-specified minimum support (minsup).
2. Generation of confident association rules from the found frequent itemsets. In other words, a rule with a confidence value above a user-specified minimum confidence (minconf).

2.2 Evolution of the Concept

Association rule mining is a datamining technique characterized by its understandability, simplicity over other supervised techniques, intuitiveness and, ease of implementation. This technique represents a simple method, free from model-based assumptions, that is, it eschews linearity assumptions underlying many classical supervised classification, regression, and ranking methods, and directly models conditional probabilities (ex: $P(y|a)$). In fact, association rule mining proceeds by looking for correlations based on subsets of frequently co-occurring past events (items) in order to generate a set of if-then rules. Unlike classification rule mining techniques which can predict only one attribute, association rules involve the prediction of any attribute in the data set. The generated rules form, then, richer models that are, moreover, interpretable and easy to understand by human users (i.e., given the rule if X then Y, it is obvious that Y was recommended because X is satisfied).

The key strength of ARM is its completeness and exhaustivity in terms of generation of rules. Indeed, in traditional classification techniques (e.g., decision tree, decision lists, neural networks, etc.) a small subset of rules is produced, based on various heuristics, however, ARM consists in finding all the frequently appearing patterns in the given database. Thus it doesn't miss any detailed rule that might be important in some application cases. Another main advantage of ARM over some other supervised learning paradigms is its ability to handle the "cold start" problem (i.e., related to small training datasets). Actually, in these paradigms, a large sample is necessary for generalization, as bounds scale only with the sample size. However, in association rules, the user-specified minsup threshold guarantees that predictions can be made only when there are enough data, even if it is a small amount of it.

Besides to their proper strengths ARM has, over years, undergone improvements and extensions. Althought its capability to generate rules in linear time and even scale up to large database, first algorithms (i.e, Apriori) presented three main challenges: the multiple scans of transaction database, the generation of a huge number of candidates, and the tedious workload of support counting for candidates.Tremendous number of algorithms attempting to deal with these challenges and aiming at improving the computational capacity and the use efficiency of the memory have been reported. Most of them have focused on three main ideas:

- Reducing passes of transaction database scans like partitioning-based algorithms, transaction reducing algorithms, and algorithms without candidate generation (FP-growth, H-mine), etc.)
- Shrinking the number of candidates: sampling-based algorithms, hash-based algorithms, condition-based algorithms, etc.
- Facilitating support counting of candidates like algorithms using vertical data format (ECLAT), hash-based algorithms, etc.

More explanation of these latter with references of examples of their implementations and examples of algorithms corresponding to further methodologies, can be found in (Aggarwal et al., 2014).

Actually, the numerous efforts made to improve the performance of existing mining algorithms, has, considerably, helped with the objective of efficiently handling the huge amount of data required for some prediction applications. Besides to these efforts mainly related to performance issues, other advances in association rule mining have been proposed which enhanced their analytical and most importantly predictive capabilities and broaden their application scope (e.g., spatio-temporal dynamics). Among these one may mention, spatio-temporal association rule mining to capture time and space dependent patterns, quantitative and fuzzy association rule mining to better handle real life descriptors, mining with multiple minsups to not skip rarely occurring but important patterns, and class association rule mining to cover other tasks such as classification and prediction. Class Association rule (CAR) mining algorithms consist in discovering rules in the form of $a \longrightarrow b$, where the consequence part (b) have to be an item labelled as a class. These rules can then be used to predict the class of unclassified records. This method enables users to decrease the number of generated rules by precising which form of rule they are interested in. CARs showed, as well, success in different prediction applications. It was even reported in many experimental studies, as for instance (Thabtah et al., 2004), that this approach can outperform traditional classification methods, such as decision trees, in constructing a more accurate predictive systems.

2.3 Association Rules in the Prediction Framework

A predictive association-based model uses the antecedent of a rule to predict the consequent of the rule. In other words, it indicates what item is likely to occur given the occurrence of a certain itemset. This type of models have been employed for many applications in different domains, one may cite predicting customer's likely future purchases, in the context of basket market analysis (Chen et al., 2014); in biology, the prediction of proteins functions, based on the proteinprotein interaction networks (Park et al., 2015); in medicine, the prediction of the risk level of the patients having heart disease (Ilayaraja and Meyyappan, 2015); and last but not least, in the energy management field association rules have been employed for the prediction of buildings Occupant Location. In fact, this problem consists in determining the location of buildings occupants, based on their movement historic, to maximize heating, ventilation, air conditioning (HCAC) energy efficiency. In other words, operate HCAC systems in accordance with occupant movements to satisfy their needs without squandering energy (Ryan and Brown, 2013).

Several other works such as (Lin and Li, 2015) have demonstrated how association rule mining can also be employed for extracting spatio-temporal patterns related, for instance, to urban dynamics (e.g., urban growth or land use/cover change). This type of application discovers patterns related to spatio-temporal relationships which are typically embedded in geospatial data instead of being explicitly encoded in the database. Spatial association rules handle pieces of information such as location and topology of items to extract frequent patterns mainly showing the interaction of two or more space-depending attributes or spatial objects. While the Temporal rules, captures timedependent attributes to reveal knowledge about the cyclic, periodic or sequential nature of some patterns. The real added value of mining association rules in a temporal context is to gain more predictive capabilities. For instance in sequential rule mining (Rudin et al., 2013), both, the occurrences of items and the order between these occurrences counts in producing rules that attempt to determine what event will next be revealed based on sequences of past events.

Spatial and temporal relationships could have hierarchical aspects. For instance, in the case of land cover an industrial zone can be also denoted as a developed zone in a higher hierarchy level. Hence both rules involving ”developed zones” and others involving ”industrial zones” are then generated. In this context, Association rules has been extended to Multilevel association rules which support the hierarchical aspect of patterns such as spatio-temporal relationships. Association rules are conventionally designed for handling only categorical data. However, spatial relationships include as well, metric relationships such as distance that are described with numerical data. This issue has been addressed by presenting quantitative association rule mining which has been combined afterward with fuzzy sets theories to address the imperfect aspects of spatial relationships (e.g., determining degrees of neighborhood through partial membership to intervals of the distance attribute) (Farzanyar and Kangavari, 2012).

3 OUR PROPOSAL

In this section we propose an apriori based approach for mining rules predicting the evolution of geographical areas. Our goal is to demonstrate how the advances mentioned in section 2 can be harnessed to make association rule mining suitable for this prediction problem. Based on spatio-temporal data, we aim at extracting rules which involve particular temporal and spatial relationships in order to indicate the next evolution (i.e., next land cover for a certain geographical zone).

In the present work, geographical entities are characterized by their functions (e.g., land use/cover), spatial relationships are represented by their neighborhood interrelations, and temporal relationships are represented by sequential occurrences (successions) of GEs (see figure 1).

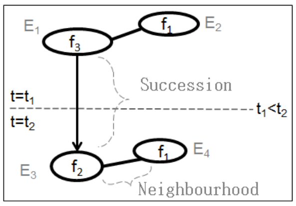

Figure 1: The Considered Spatio-temporal Relationships.

As regards identifying the spatial relationships of neighbourhood among entities, two approaches may be adopted. The first one defines neighbouring entities as directly adjacent ones, and the second, considerate that, given two GEs E1 and E2, E2 is a neighbour of E1 if the distance between them is under or equal to a user-specified threshold. In the second case, the continuous numerical attribute of distance will be discretized into two categorical values: neighbour and non-neighbour. However, the strict membership or non-membership to either of these categories can lead to a problem of misestimating distance values near the borders. In order to deal with this problem, we propose to consider the concept of partial membership using a membership function defined based on fuzzy sets theories.

It is worth noting that this approach supports two different applications: addressing entities consisting in geographically delimited zones and addressing buildings as the studied entities. In the first case, the functional level of evolution is related to land use/cover change, and in the second, it regards functions of building (e.g., residential, industrial, administrative, leisure equipements, etc)

3.1 Dataset

In our context, the Spatio-temporal database describes all the geographical entities situated in the studied geographical zones at different consecutive dates. Corine Land Cover Database is an example, of which an extract is given in figure 5. It provides pieces of information regarding their functional (e.g., land cover), and spatial characteristics (e.g., locations, surface area, form, etc).

The first task in applying association rule mining consists in proposing, from the available spatio-

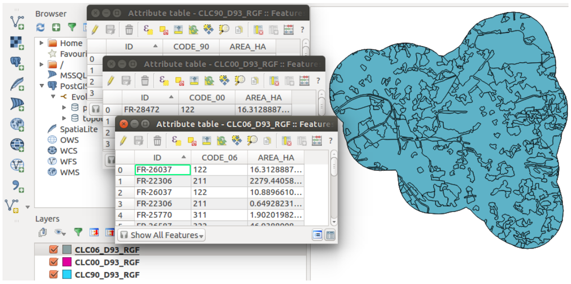

Figure 2: Spatio-temporal Data.

temporal data, an appropriate format for the learning dataset. This latter will, afterwards, be used to extract adequate rules for our prediction problem. The dataset, in the present approach, is composed of instances that each represents the evolution history of a geographical entity (see figure 3.a).

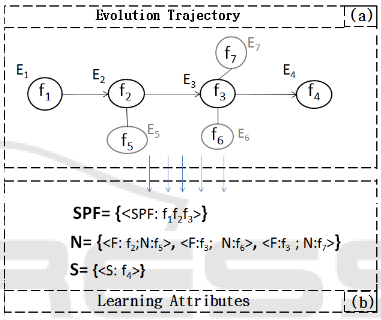

Figure 3: An Example Illustrating the Extraction of Learning Attributes.

In fact, the evolution trajectory (history) of a certain geographical entity is encoded in the form of a transaction composed by:

- A sequence of its past evolutions (sequence of past functions). In the example illustrated in figure 3.b the SPF attribute corresponds to $< SPF : f_1 f_2 f_3 >$.
- A set of items representing the functions of the neighbouring entities of each entity involved in the evolution sequence . In the example above, $< F : f_2$;**N:**$f_5 >$, represents the neighbourhood relationship between the entities E2 and E5, respectively, characterized by the functions f_2 and f_5.
- An item representing the succeeding function (f_4)

Hence, the set of attributes of our dataset is A={spf_1 , spf_2, spf_3 ,..., spf_n, N_1,N_2,..., N_x, s_1, s_2,..., s_n}. n denotes both the number of transactions which is equal to the number of sequences of past evolutions, and the number of successors. x denotes the number of neighbours, n<F={f_1, f_2, f_3, ..., f_x} with F is the set of all possible functions of entities. In the context of the

Attributes

Instances	$<F:f_1>$	$<F:f_2>$	$<F:f_3>$	$<F:f_4>$	$<SPF:f_1f_2f_3>$	$<F:f_2;N:f_5>$	$<F:f_3;N:f_6>$	$<F:f_3;N:f_7>$	$<S:f_4>$	...
Seq1	1	1	1	1	1	1	1	1	1	...
Seq2	1	0	1	0	0	0	1	1	0	...
Seq3										...
...										

Figure 4: Structure of the Learning Database.

example above, 4 depicts the structure of the learning database. Rows correspond to the different evolution trajectories and columns to the attributes (SPF, N, S).

Using this dataset format, our objective is to find rules that predict the attribute *S* based on the other attributes *N* and *SPF*. A predictive evolution rule should be in this form:

$$< SPF =< f_1 f_2 f_3 > \wedge < F = f_2; N = f_5 > \longrightarrow < S = f_4 >$$

3.2 Preliminary Results

```
Best rules found:
 1. <F:112;N:311>=1 <F:112;N:231>=1 316 ==> <F:112;N:121>=1 316    conf:(1)
 2. <F:112;N:211>=1 <F:112;N:311>=1 <F:112;N:231>=1 315 ==> <F:112;N:121>=1 315    conf:(1)
 3. <F:112;N:141>=1 <F:112;N:111>=1 313 ==> <F:112;N:121>=1 313    conf:(1)
 4. <F:112;N:141>=1 <F:112;N:142>=1 313 ==> <F:112;N:121>=1 313    conf:(1)
 5. <F:112;N:111>=1 <F:112;N:142>=1 313 ==> <F:112;N:121>=1 313    conf:(1)
 6. <F:112;N:141>=1 <F:112;N:111>=1 <F:112;N:142>=1 312 ==> <F:112;N:121>=1 312    conf:(1)
 7. <F:112;N:311>=1 <F:112;N:122>=1 311 ==> <F:112;N:211>=1 311    conf:(1)
 8. <F:112;N:141>=1 <F:112;N:511>=1 310 ==> <F:112;N:121>=1 310    conf:(1)
 9. <F:112;N:142>=1 <F:112;N:511>=1 310 ==> <F:112;N:121>=1 310    conf:(1)
10. <F:112;N:311>=1 <F:112;N:242>=1 310 ==> <F:112;N:211>=1 310    conf:(1)
```

Figure 5: An extract of First Generated Rules.

In order to test our proposed dataset format (figure 4) mainly in terms of relevance of generated rules, we run Apriori. The preliminary results showed that only rules involving the neighbourhood attribute have been generated. Besides their form does not match the specified interesting form which requires solutions related to: considering rare but important items (SPF, S), and guaranteeing the generation of interesting rules only (i.e., $SPF \wedge N \longrightarrow S$).

3.3 Generation of Candidates

As mentioned above, apriori-based algorithms proceeds on two steps: the first one consists in generating candidate itemsets then keeping only the frequent ones, and the second step consists in generating rules from the found frequent itemsets. For each frequent itemset f apriori finds all their nonempty subsets α then generate confident rules (i.e., whose confidence is above minconf threshold) of the form:

$$(f - \alpha) \longrightarrow \alpha$$

For instance, given the frequent itemset f=$\{A,B,C\}$, α=$\{A, B, C, AB, AC, BC$. An example of generated rules is $B \wedge C \longrightarrow A$. Due to this, a great importance is granted to the candidate generation step, in order to prevent generating useless candidates and consequently useless rules. Actually, in our approach we propose to predict the successor attribute (*S*) based on the other attributes (i.e., *SPF* and *N*). Hence, any itemset which misses any of these attributes (*S*, *SPF*, *N*) cannot produce rules in the required format specified above. Therefore we propose in our frequent pattern generation function (*FP-Gen( )*) to add a constraint in order to only keep candidates which include at least one item for each attribute (see the example in figure 6).

Candidat Itemsets	Evaluation
$\{<F:f_2;N:f_5>, <SPF: f_1,f_2,f_3>, <S:f_4>\}$	✔
$\{<F:_3;N:f_7>, <F:f_3;N:f_6>\}$	✖
$\{<SPF: f_1,f_2,f_3>, <S:f_4>\}$	✖

Figure 6: Validation of candidat itemsets.

As regards to the *SPF* attribute, our *FP-Gen* function has, also, to handle its sequential nature when generating candidates as in this attribute, not only the occurrence of entities matter, but also their order. The key element in this first step of apriori (the frequent itemsets generation) is the minsup threshold (*minsup*) as it is used to assess the frequency of candidate itemsets. However, using a single minsup assumes that all items have similar frequencies in the dataset, which is not the case for different real-life application and indeed for our prediction problem. Actually, in our dataset, neighbourhood items are a way more frequent than other items and setting the *minsup* too low to capture rare items can lead to combinatorial explosion. Therefore we propose, as a solution, to specify multiple minsups, in other words, set a different *minsup* for each attribute i.e., *minsup(N)* $\neq$ *minsup(SPF)* $\neq$ *minsup(S)*. The key element in this first step of apriori (the frequent itemsets generation) is the minsup threshold (*minsup*) as it is used to assess the frequency of candidate itemsets. However, using a single minsup assumes that all items have similar frequencies in the dataset, which is not the case for different real-life application and indeed for our prediction problem. Actually, in our dataset, neighbourhood items are a way more frequent than other items and setting the *minsup* too low to capture rare items can lead to combinatorial explosion. Therefore we propose, as a solution, to specify multiple minsups, in other words, set a different *minsup* for each attribute i.e., *minsup(N)* $\neq$ *minsup(SPF)* $\neq$ *minsup(S)*.

3.4 Generation of Predictive Rules

With regards to rule generation, we aim at generating only rules where S attributes are in the consequent part of the rule and the other attributes are in the antecedent part. In this regards, we suggest the use of class association rule mining wherein each transaction is labelled with a class *s*. Let *S* be the set of all class items corresponding to the attribute *S*, I be the set of all items in database corresponding to the attributes *SPF* and *N*, and $S \cap I = \emptyset$. A class association rule is an implication in the form: $a \longrightarrow b$, where $a \subseteq I$ and $b \subseteq S$. In *CARs* candidats are denoted as **ruleitems** and are in the following form: *(Condset,s)*. *condset* represents a set of items (condset $\subseteq$ I) and *s* a class (s $\subseteq$ S). Each ruleitem represents the rule *condset* $\longrightarrow$ *s*. The support count of a *condset* (*SCC*) represents the number of transactions in the database which contain the *condset*. The support count of the generated rules (*SCR*) is the number of transactions of the database containing the *condset* and having *s* as a class. Like in the traditional association rule mining, apriori in CARM generates all frequent *ruleitems* whose support is above a minsup threshold, then, use them to generate class association rules with a confidence value (*SCR* / *SCC*) above the user-specified minconf threshold. The distinct particularity in CAR candidate generation function is that, the joining step is done by joining condsets of ruleitems having the same class.

As mentioned above this approach (CARM) aims at only producing rules in a specific form which is assumed to be adequate for our prediction task. Although it may seem logical to proceed, simply, to a post-selection of interesting rules, this solution is, in practice, very difficult and sometimes impossible due to the combinatorial explosion, in other words, the huge number of rules that could be generated.

4 CONCLUSION

Association rule mining is a fundamental datamining task which has been basically presented as an exploratory tool rather than a predictive tool. In this research we attempted at reviewing the most important advances, in this datamining tool, ranging from proposing scalable algorithms and efficient methodologies for mining frequent itemsets to handling a diversity of data types and structures and extended mining tasks. Our overall goal is to show how all this progress has contributed in making AR a powerful prediction tool. Indeed, we proposed an association rule-based approach for predicting the evolution of geographical areas as an attempt to show how the progress, done so far, can, practically, be harnessed for this example of prediction problem.

Our proposal consists in an apriori-based approach for mining rules predicting the evolevolution of geographical areas. This approach proposes to address issues related to handling spatial and temporal relationships in the learning dataset, producing rules involving rare patterns, and making sure to only generate rules in an adequate form for our prediction problem.

REFERENCES

Aggarwal, C. C., Bhuiyan, M. A., and Hasan, M. A. (2014). Frequent pattern mining algorithms: A survey. In Aggarwal, C. C. and Han, J., editors, *Frequent Pattern Mining*, pages 19–64. Springer International Publishing.

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM.

Chen, J., Miller, C., and Dagher, G. (2014). Product recommendation system for small online retailers using association rules mining. In *Innovative Design and Manufacturing (ICIDM), Proceedings of the 2014 International Conference on*, pages 71–77.

Farzanyar, Z. and Kangavari, M. R. (2012). Efficient mining of fuzzy association rules from the pre-processed dataset. *Computing and Informatics*, 31(2):331–347.

Gharbi, A., de Runz, C., Faiz, S., and Akdag, H. (2014). An association rules based approach to predict semantic land use evolution in the french city of saint-denis. *International Journal of Data Warehousing and Mining*, 10:1–17.

Ilayaraja, M. and Meyyappan, T. (2015). Efficient data mining method to predict the risk of heart diseases through frequent itemsets. *Procedia Computer Science*, 70:586 – 592. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems.

Lin, J. and Li, X. (2015). Knowledge transfer for large-scale urban growth modeling based on formal concept analysis. *Transactions in GIS*, pages n/a–n/a.

Park, H. A., Kim, T., Li, M., Shon, H. S., Park, J. S., and Ryu, K. H. (2015). Application of gap-constraints given sequential frequent pattern mining for protein function prediction. *Osong Public Health and Research Perspectives*, 6(2):112 – 120.

Rudin, C., Letham, B., and Madigan, D. (2013). Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, 14:3441–3492.

Ryan, C. and Brown, K. (2013). Predicting occupant locations using association rule mining. In Bramer, M. and Petridis, M., editors, *Research and Development in Intelligent Systems XXX*, pages 63–77. Springer International Publishing.

Thabtah, F., Cowling, P., and Peng, Y. (2004). Mmac: a new multi-class, multi-label associative classification approach. In *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, pages 217–224.