International Journal of Pure and Applied Mathematics
Volume 119 No. 16 2018, 5265-5279
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

CBIR: Classification based Association Rules and Approaches in Data Mining

[1]P. Pavankumar and [2]Rashmi Agarwal
[1]Computer Science and Engineering,
Madhav University,
Abu Road, Sirohi, Rajasthan.
pavankumar.palla@gmail.com
[2]Computer Science and Engineering,
Madhav University,
Abu Road, Sirohi, Rajasthan.
rashmiagarwalcse@gmail.com

Abstract

Most existing text mining methods adopted term-based approaches but they all suffer from the problems of polysemy and synonymy. Polysemy is the word which giving the multiple meaning of word and synonymy is the word which giving the similar meaning of word. After some years, people have been adopted pattern based approaches should perform better than the term-based approaches. This paper with proposed system implements innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information with effective patterns as per the users requirement. Here user also gets the meaningful information without mis interpretation problem. This research deals with several classifiers including gk-Nearest Neighbor(k-NN), Radial Basis Function(RBF), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) which are used as trained classifiers for per- forming classification of data into relevant and non-relevant data also uses ARM(Association Rule Mining) with CBA (Classification bases Association Rules). This study intends to compare the efficiency of the various existing classification algorithms with the proposed classification algorithms on the basis of run time, error rate and accuracy

Key Words: Association Rule Mining (ARM), Apriori, CBA, UCI Machine Learning Repository.

1. Introduction

Data and data have been acknowledged as a profitable resource since long time. In any case, the use of data and the apparatuses for utilizing that data has been changed a considerable measure after some time. During 1960's database creations were and more network popular, the relational DBM. In database display and social DBMS usage came into utilization Propelled database.

Association Rule Mining

Association rule mining is one of the essential and all around investigated techniques of data mining to discover vital connections among data things. In light of the design of database, 28 distinct techniques have been created formining the data. The flat design of database is utilized by Apriori arrangement strategies while vertical format is the base of FP-development and Éclat calculations. A different methodology either improves the productivity of the current methodologies or manages abnormal state data mining ideas.

The most agreeable order of data mining techniques is on the premise of the design of the database under thought. Diverse methodologies have been suggested that utilization even design of database, vertical format of database or anticipated format of database. A few scientists deal with enhancing the productivity of the mining process while others attempted to uncover progressed, confused and abnormal state information from the database. Additionally, swarm insight techniques have been utilized as a part of various fields for different assignments going from advancement to appropriation of assets. The utilization of swarm insight for data mining has turned out to be well known since most recent two decades. After that few developments in the field of data mining utilizing swarm knowledge has been completed. This section contemplates light on the accessible writing in both the field's viz. data mining and swarm knowledge and likewise introduces a talk of the fruitful applications of various swarm insight techniques in data mining.

Evolution of Data Mining Techniques

Data and data have been acknowledged as a profitable resource since long time. In any case, the use of data and the apparatuses for utilizing that data has been changed a considerable measure after some time. During 1960's data b Data and data have been acknowledged as a profitable resource since long time. In any case, the use of data and the apparatuses for utilizing that data has been changed a considerable measure after some time. During 1960's database creations were and more network popular, the relational DBM. In database display and social DBMS usage came into utilization. Propelled databases are creations were and more network popular, the relational DBM. In database display and social DBMS usage came into utilization. Propelled database.

Association Rule Mining

Association rule mining is one of the essential and all around investigated

techniques of data mining to discover vital connections among data things.

In light of the design of database, 28 distinct techniques have been created for mining the data. The flat design of database is utilized by Apriori arrangement strategies while vertical format is the base of FP-development and Éclat calculations. A different methodology either improves the productivity of the current methodologies or manages abnormal state data mining ideas. These all around preferred methodologies are talked about below:

Techniques based on Horizontal Layout of Databases

The principal calculation to produce all continuous itemsets was proposed by Agrawal et al. [AGR1993] and named AIS (after the name of its proposers Agrawal, Imielinski and Swami). The calculation produces all the conceivable itemsets at each level of traversal. In this way, it produces and stores visit and occasional itemsets in each pass. Era of occasional itemsets was undesirable and was a noteworthy downside over its execution. Later on, AIS was enhanced and renamed as Apriori by Agrawal et al. The new calculation utilizes a level-wise and broadness initially looks for generating association rules. Apriori and Apriori Tid calculations create the candidate itemsets by utilizing just the itemsets discovered vast in the past pass and without utilizing the value-based database. Apriori utilizes the descending conclusion property of the itemset support to prune the itemset grid the property that all subsets of incessant itemsets must themselves be visit.

A comparable calculation called Dynamic Itemset Counting (DIC) was proposed by Brin et al. in [BRI1997]. DIC parcels a database into a few squares set apart by begin focuses and more than once checks the database. Not at all like Apriori, can DIC include new candidate itemsets at any begin point, rather than exactly toward the start of new database check. At each begin point, DIC gauges the help of all itemsets that are right now numbered and add new itemsets to the set if every one of its subsets is evaluated to be visit.

2. Existing Work

Association Rule Mining
Notations and Basic Concepts

Association rule mining is the technique of finding association rules that satisfy the predefined minimum support and confidence from a given database. This technique is widely adopted in the market basket analysis and currently used in various fields where relativity of the attributes plays a vital role in deciding the functionality of respective domains. Association rule is a relation between a pair of disjoint item sets. If LHS and RHS are two disjoint sets of items, the association rule is stated as LHS→ RHS. LHS and RHS are sets of items, the RHS set being likely to occur whenever the LHS set occurs

Formal Problem Description

Let I = $\{i_1, i_2 \ldots i_n\}$ be a set of items. Let D be a set of transactions or database.

Each transaction t∈D is an item set such that t is a proper subset of I. A transaction t supports X, a set of items in I, if X is a proper subset of t. An association rule is an implication of the form X→Y, where X and Y are subsets of I and X∩Y= Ø. X is called the antecedent and Y is called the consequent part. The support of rule X→Y denotes ratio of the number of transactions in the database that contains the itemset X and Y to total number of the transactions in the database D. The confidence of rule is the ratio of the number of transactions in the database that contains the itemset X and Y to number of the transactions that contains X. A rule X→Y is strong if it reaches the minimum support threshold and minimum confidence threshold. Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve only those rules having support and confidence higher than the user specified minimum support and confidence threshold [MIN1996]. Association rule mining consists of two stages viz. the discovery of frequent itemsets and the generation of association rules. It follows that in majority of cases, the discovery of the frequent set dominates the performance of the whole process.

Frequent Item set Generation

Visit thing sets are those arrangements of things whose events surpass a predefined edge in the database. The computational prerequisites for visit item set generation are generally more costly than those of run generation. Generating all the subsets of set of things, say I {i1, i2… in} for large value n, is practically a lot of troublesome because of the gigantic search space. In fact a linearly developing number of things suggest an exponential developing number of item sets [ZAK1997]. The way toward generating incessant item sets can be additionally isolated into two sub-issues: Candidate large item sets generation handle and successive item sets generation prepare. Item sets that are normal or have the plan to be large or incessant are called candidate item sets and among those the item sets whose help surpasses the help edge are viewed as regular item sets.

A beast drive approach for finding continuous item sets is to decide the help mean each candidate item set. For this reason, a counter for each item set may be utilized and initialized to zero. At that point all transactions are scanned and at whatever point one of the candidates is perceived as a subset of a transaction, its counter is augmented [AGR1994]. Another approach is to decide the help values by set crossing points. Each thing may be associated with a tidlist, which is the rundown of transactions that contain the thing. Each transaction is indicated by special transaction identifier (tid). Accordingly, there will be a tidlist associated with each item set X and to obtain the tidlist of a candidate item set U=X ∪ Y, we calculate X.tidlist ∩ Y.tidlist. Finally the actual help is generated by deciding |U.tidlist|. Whichever approach is utilized, the computational unpredictability of the procedure is always remarkable. There are many approaches that can diminish this intricacy and enhance the performance of the procedure. Apriori standard is one such approach proposed by

[AGR1993]. The Apriori rule states that if an item set is visit then all of its subsets should also be visit. On the other hand, if an item set is occasional then all of its supersets must be rare as well. This may help in lessening the quantity of candidate item sets investigated as the subsets of continuous item sets and supersets of rare item sets require not to be investigated separately.

Classification based Association Rules (CBA)

The rules coming about because of Associative Classification mining can be assessed to choose a subset of the rules that will frame the model or classifier. To the best of our insight, Liu, Hsu, and Ma were the rest to create a classifier in light of affiliation rules. They demonstrate that the classifier constructed executes and also or superior to anything surely understood decision tree calculations. From that point forward, numerous affiliation administer based classifiers have been worked for different areas. Among others, for classifying mammography pictures, for classifying web documents, for re-commender frameworks, for classifying spatial information, for document classification, and for content arrangement. The way toward building the classifier includes choosing rules by certainty or support. Certainty is a well known standard for run determination to the classifier as it signifies the strength of a run the show. On account of CBA, they utilize a heuristic to choose a subset of the rules that orders the preparation set generally precisely. At times, the pruning is as basic as evacuating contradicting rules or more confused like utilizing post pruning techniques that are utilized as a part of decision trees. In CBA-CB, the created CARs are requested in light of the accompanying definition.

Definition 2.2.1 Rule Ordering Association

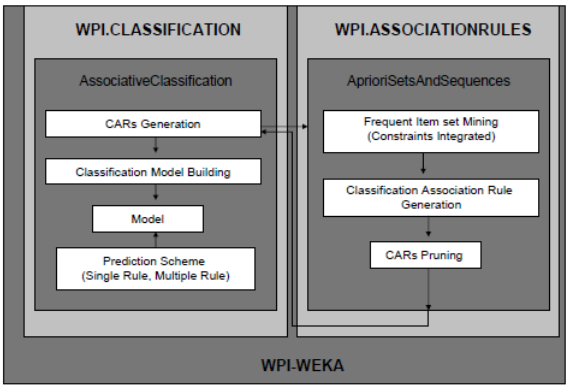

Figure 3.1: Architecture of WPI Classification System

Given two rules, ri and rj , ri, rj (ri goes before rj) if the certainty of ri is more noteworthy than that of rj or, their certainty are the same, however the help of ri is more prominent than that of rj , or, both the certainty and the help of ri and rj are the same, yet ri is produced sooner than rj . Give R a chance to be the arrangement of CARs and D be the preparation information. The point of the model development algorithm is to pick an arrangement of profoundly

predictive rules in R to cover the preparation information D. The classifier constructed is of the accompanying structure: < r_1; r_2 ; :::; r_n; default class > where r_i 2 R, r_a r_b if a < b. Default class is the default mark utilized when none of the rules can classify a case. Algorithm 2 demonstrates the CBA-CB technique. In stage 1, the rules are sorted by the request said above; at that point each administer is considered thusly.

Algorithm 2 CBA-CB Algorithm
Inputs: rules R, training set instances D
Output: classifier C

1. R = sort(R);
2. for each rule r 2 R in sequence do
3. temp = ;
4. for each instance d 2 D do
5. if d satisfies the conditions of r then
6. store d.id in temp and mark r if it correctly classifies d;
7. end if
8. end for
9. if r is marked then
10. insert r at the end of C ;
11. delete all the cases with the ids in temp from D;
12. select the default class for the current C ;
13. compute the total number of errors of C ;
14. end if
15. end for
16. Find the rst rule p in C such that C_p, the list of rules in C up to p, has the lowest total number of errors. and drop all the rules.
17. Add the default class associated with p to the end of C, and return C

3. Proposed Work

Implementation

We have implemented our classification framework in WEKA. WEKA is an open-source suite of machine learning algorithms. The inspiration for actualizing our theory in WEKA is the broad utilization of this framework in WPI's Knowledge Discovery and Data Mining Research Group. WEKA is created in the Java Programming Language. Figure 4.1 demonstrates the engineering of our classification framework. We modified the current Apriori like algorithm, Apriori Sets and Sequences, to generate classification association rules. The generated rules are utilized for building models. The resulting models are tried for accuracy.

Alluding to Figure 4.1, the association rule based classification algorithm is called Associative Classification and is a piece of the Wpi. Classifiers bundle. We demonstrate the communication between Associative Classification and Apriori Sets and Sequences. We additionally demonstrate the distinctive modules in both the algorithms.

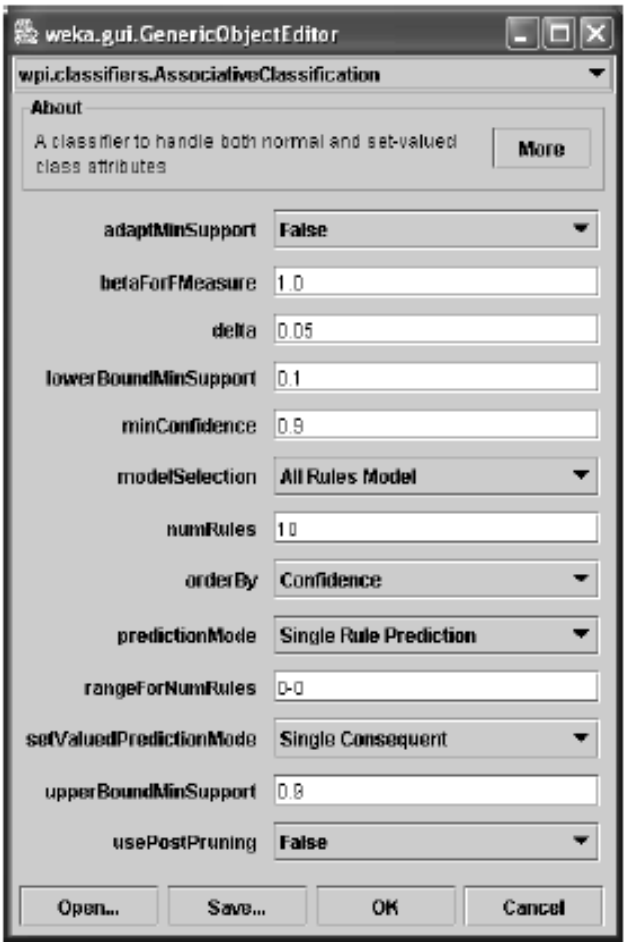

Figure 3.2: Parameter Menu for Our Extended Association Rule Mining

Algorithm 3 is the modified control procedure to mine (compelled) CARS. In this proposal, we have modified the first control procedure (see Algorithm 1) to permit pruning of rules in light of pessimistic error. We have additionally modified the algorithm to take into account nearness or nonappearance of items (semantic constraints) in the rules. All the more accurately, clients can specify an item to show up or not to show up on either the forerunner or the subsequent of a rule. We utilize a pruning technique to generate just item exhausts, threshold, sets that can potentially turn out to be a piece of the client indicated rules. In the rule generation advance, before a rule is generated, it is verified whether it fulfills the client determined constraints and, provided that this is true, the rule gets generated. WEKA contains numerous outstanding classification algorithms and one noteworthy commitment of this proposition is the classifier in light of Association rule mining algorithm.

Info parameters incorporate required Antecedent, required Consequent, dis-allowed Antecedent and dis-allowed Consequent. The while circle in Step 5 rehashes itself until the point when the support limit is beneath the minsupport or the quantity of rules generated success as per the client indicated number of rules. In the event that we investigate the iterative procedure of generating item sets and rules from them: In Step 6, we generate the 1-item item sets. In stages 11-15, the condition debilitates all conceivable item sets that can be created until the point when no more items of size k can be joined to deliver items of size (k+1). Just those item sets that will potentially yield rules with the required item sets are generated. The algorithm for this can be found in Section 4.3. The

Algorithm 3 Modified Apriori Sets and Sequences Control Procedure

Inputs: required Antecedents, required Consequents, disallowed Antecedents, disallowed Consequents, num Rules

Outputs: rules

1. rules = ;
2. support = UpperBoundSupport;
3. freqItemsets = ;;
4. requiredItems = requiredAntecedents [requiredConsequents
5. while (support > minsupport AND rules.size < numRules) do
6. L_1 = f1-item itemsetsg;
7. for (k = 2; L_{k1} 6= ;) do
8. C_k = generate Candidates(L_{k1} , required Items);
9. L_k = evaluate Candidates(C_k);
10. Freq Itemsets [L(k);
11. end for
12. maxFreqItemsets = genMaxFreqItemset(freqItemsets);
13. rules = Generate All Rules(maxFreqItemsets, requiredAntecedents, requiredConse-quents)
14. rules = PruneRules(rules);
15. if (rules.size > minRules) then
16. return rules;
17. end if
18. support = support - delta;
19. end while

Experimental Evaluation

Evaluation Metrics

We assess the classifier in view of error rate with different prediction plans. We additionally report the accuracy rate. The error rate signifies the quantity of wrong predictions over the aggregate number of predictions. The accuracy rate means the quantity of right predictions over the aggregate number of predictions.

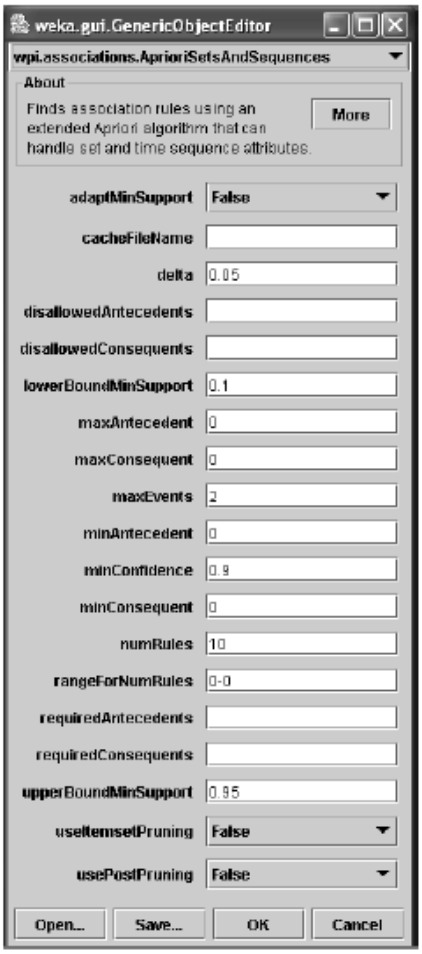

Figure 3.1: Parameter Menu for Our Extended Association Rule Mining

Experimental Results

We divide this section into two parts. In part 1, we focus on the improvements made to Apriori Sets And Sequences by item set pruning in the presence of constraints. Here we evaluate performance based on time taken for mining and generating rules and the number of maximal frequent item sets generated. A frequent item set is considered maximally frequent if none of its supersets is frequent.

Data Set

We tested the classification system with the following datasets obtained from the UCI Machine Learning Repository: census-income, mushroom and forest cover. Table 4.5 shows the properties of these datasets. As part of pre-processing, continuous valued attributes were discretized using WEKA's instance based discretizationlter with the number of bins set to 10.

Table 3.1: Dataset Properties

Dataset	# attr	class	# class values	# instances
sonar	61	rocks/mines	2	208
census-income	15	income-level	2	32,561
mushroom	23	edible/poisonous	2	8,124
forest cover	17	forest cover type	7	74,056

Item set Pruning in the Presence of Constraints

As a major aspect of our experiments, we were interested in contrasting thing set pruning versus non-pruning. We ran experiments with the mushroom, evaluation pay and backwoods cover data sets. We produced single and multiple constraint classification rules. We watched the subsequent parameters, for example, the quantity of thing sets delivered, number of maximal thing sets created and time taken for generating rules.

Table 3.1 demonstrates the parameters utilized as a part of running the experiments. In these experiments, the goal was to create whatever number standards as could be allowed with the help more prominent than or equivalent to 1%. The base confidence was set to half.

Table 3.2: Experimental Parameters

Prune	Req. Ant	Req. Con	itemsets	Rules	Max. itemsets	Time(s)
No	none	Class	45391	21101	158	4951
Yes	none	Class	42620	21101	42	4357
No	odor	Class	45391	8288	158	1153
Yes	odor	Class	33160	8288	26	1813

Table 3.3: Comparison of Constraint-based Pruning vs. Non-Pruning for Mushroom Dataset

Prune	Req. Ant	Req. Con	itemsets	Rules	Max. itemsets	Time(s)
No	none	class	45391	21101	158	4951
Yes	none	class	42620	21101	42	4357
No	odor	class	45391	8288	158	1153
Yes	odor	class	33160	8288	26	1813

In Table 3.3, we show the results for the mushroom data set. The rest column appears if constraint based pruning was chosen or not. On account of pruning being exchanged o, all competitor thing sets are utilized as a part of generating valid thing sets at each level of the Apriori procedure.

In looking at the rest two columns (single constraint), we watch the lessening in the quantity of thing sets delivered and the decrease in time taken for generating the guidelines.

Be that as it may, interestingly, in the following two columns (twofold constraint) despite the fact that the quantity of thing sets delivered diminishes, the time taken increments.

We gured this is where countless subsets are dropped from consideration because of the constraint based pruning. The nal output of the database for help of those things costs a huge time, expanding the general time.

Table 3.4: Comparison of Constraint-based Pruning vs. Non-Pruning for Census-Income Dataset

Prune	Req. Ant	Req. Con	Item sets	Rules	Max. item sets	Time(s)
No	none	class	1071	350	82	85
Yes	none	class	410	350	36	88
No	relationship	class	1071	22	82	87
Yes	relationship	class	100	22	31	16

As seen in Table 3.3, in the case of a single constraint, the results for pruning and non-pruning are very similar. In the case of two constraints, the pruning leads to better performance in terms of time, approximately 1/5 of the time taken without pruning.

Table 3.5: Comparison of Constraint-based Pruning vs. Non-Pruning for Forest-Cover Dataset

Prune	Req. Ant	Req. Con	Item sets	Rules	Max. item sets	Time(s)
No	none	class	4297	1247	45	651
Yes	none	class	2673	1247	19	613
No	aspect	class	4297	144	45	678
Yes	aspect	class	605	144	22	208

4. Conclusions

We developed a technique to incorporate client defined information constraints during the mining of association rules. We considered two kinds of constraints to be specific semantic and synctatic. Semantic constraints require the nearness/nonattendance of specific item sets in the rules; while synctatic constraints restrain the quantity of trait esteem combines in either the antecedent or the consequent of a rule. We developed a characterization of those thing sets that will conceivably frame rules that fulfill the given constraints. This characterization enables us to lter out from consideration all the item sets with the end goal that neither they nor any of their supersets will shape valid rules.

We developed a classification framework that is based on association rule mining in the WEKA condition. We implemented the CBA model building algorithm and compared the execution of CBA with All Rules Model (ARM) where all the mined rules are a piece of the model. We developed various modes to predict an unclassified case, for example, single rule or different rule prediction weighed by confidence/support.

Bibliography

[1] Rakesh Agrawal, Tomasz Imirlinksi, and Arun Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACMSIGMOD International Conference on Management of Data, pages 207{216,1993.

[2] Sergio A. Alvarez. Binary search strategy for adaptive minimal support. Personal Communication, 2002.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for Erik B. Sudderth, Antonio Torralba, William T. Freeman, and

Alan S. Willsky. Describing visual scenes using transformed objects and parts. International Journal of Computer Vision, 77(1–3):291–330, 2008.

[4] Josef Sivic and Andrew Zisserman. Video data mining using configurations of viewpoint invariant regions. In CVPR, pages 488–495, 2004.

[5] Fadi A. Thabtah. A review of associative classification mining. Knowledge Eng. Review, 22(1):37–65, 2007.

[6] Risivardhan Thonangi and Vikram Pudi. ACME: An associative classifier based on maximum entropy principle. In ALT, pages 122–134, 2005.

[7] H. Verlinde, Martine De Cock, and R. Boute. Fuzzy versus quantitative association rules:A fair data-driven comparison. IEEE Transactions on Systems, Man, and Cybernetics–PartB: Cybernetics, 36(3):679–683, 2005.

[8] Eibe Frank and Ian H. Witten. Data Mining. Morgan Kaufmann Publishers, 2000.

[9] Pradeep Shenoy, Jayant R. Haritsa, S. Sudarshan, Gaurav Bhalotia, Mayank Bawa, and Devavrat Shah. Turbo-charging vertical mining of large databases. In SIGMOD Conference, pages 22–33, 2000.

[10] Ashok Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In VLDB, pages 432–444, 1995.

[11] Eibe Frank and Ian H. Witten. Data Mining. Morgan Kaufmann Publishers, 2000.

[12] Pradeep Shenoy, Jayant R. Haritsa, S. Sudarshan, Gaurav Bhalotia, Mayank Bawa, and Devavrat Shah. Turbo-charging vertical mining of large databases. In SIGMOD Conference, pages 22–33, 2000.

[13] Ashok Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In VLDB, pages 432–444, 1995.

[14] William W. Cohen. Learning trees and rules with set-valued features. In Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Arti cial Intelligence Conference, pages 709{716, Menlo Park, August 4{8 1996. AAAI Press / MIT Press.

[15] Michelangelo Ceci, Annalisa Appice, and Donato Malerba. Spatial associative classification at different levels of granularity: A probabilistic approach. In PKDD, pages 99{111, 2004.

[16] Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pages 487{499. Morgan Kaufmann, 12{15 September 1994.

[17] Herve Jegou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In CVPR, pages 1169–1176, 2009.

[18] Fred´eric´ Jurie and Bill Triggs. Creating efficient codebooks for visual recognition. In ICCV, pages 604–610, 2005.

[19] Mehmet Kaya and Reda Alhajj. Integrating multi-objective genetic algorithms into clustering for fuzzy association rules mining. In Proceedings of the Fourth IEEE International Conference on Data Mining, ICDM, pages 431–434, 2004.

[20] Mehmet Kaya, Reda Alhajj, Faruk Polat, and Ahmet Arslan. Efficient automated mining of fuzzy association rules. In DEXA, pages 133–142, 2002.

[21] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In CVPR, pages 506–513, 2004.

[22] Jingen Liu, Yusuf Aytar, Omer Bilal Orhan, Jenny Han, and Mubarak Shah. University of central florida at trecvid 2007 semantic video classification and automatic search. In TRECVID, 2007.

[23] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In KDD, pages 80–86, 1998.

[24] Wenmin Li, Jiawei Han, and Jian Pei. CMAR: Accurate and efficient classification basedon multiple class-association rules. In ICDM, pages 369–376, 2001.

[25] David G. Lowe. Distinctive image features from scale-invariant key points. International Journal of Computer Vision, 60(2):91–110, 2004.

[26] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatialpyramid matching for recognizing natural scene categories. In CVPR, pages 2169–2178, 2006.