Associative Classification Method

Subjects: Computer Science, Theory & Methods | Computer Science, Artificial Intelligence

Contributor: Jamolbek Mattiev , Christopher Meza , Branko Kavsek

Machine learning techniques are ever prevalent as datasets continue to grow daily. Associative classification (AC), which combines classification and association rule mining algorithms, plays an important role in understanding big datasets that generate a large number of rules. Clustering, on the other hand, can contribute by reducing the rule space to produce compact models.

frequent itemset | class association rules | associative classification | agglomerative clustering

1. Introduction

The demand for collecting and storing substantial amounts of data is growing exponentially in every field. Extracting crucial knowledge and mining association rules from these datasets is becoming a challenge [1] due to the large amount of rules generated, causing combinatorial and coding complexity. Reducing the number of rules by pruning (selecting only useful rules) or clustering can be a good solution to tackle the aforementioned problem and play an important role in building an accurate and compact classifier (model).

Mining association rules (AR) [2] and classification rules [3][4][5][6][7][8] enable users to extract all hidden regularities from the learning dataset, which can later be used to build compact and accurate models. Another important field of data mining is associative classification (AC), which integrates association and classification rule mining fields [9]. Research studies [10][11][12][13][14][15][16][17] demonstrate that associative classification algorithms achieve better performance than “traditional” rule-based classification models on accuracy and the number of rules included in the classifier.

Clustering algorithms [18][19][20][21] group similar rules together, considering only representative rules for each cluster helps to construct compact models, especially in the case of large and dense datasets. There are two main types of clustering: partitional and hierarchical. In the case of partitional clustering [22][23], rules are grouped into disjointed clusters. In the case of hierarchical clustering [24], rules are grouped based on a nested sequence of partitions. Combining clustering and association rule mining methods [25][26] enables users to analyze accurately, explore, and identify hidden patterns in the dataset, and build compact classification models.

2. The Associative Classification Field

The “APRIORI” algorithm is executed first to create CARs in the CBA approach, which employs the vertical mining method. Then, to generate predictive rules, the algorithm applies greedy pruning based on database coverage.

CBA uses a different rule-selection procedure than the researchers do—the rule that can classify at least one training example; that is, if the body and class labels of the candidate rule match those of the training examples, the body and class labels are chosen for the classifier. Because the researchers attempted to decrease the size of the classifier, the researchers utilized clustering first and then chose the representative rule for each cluster.

The Simple Associative Classifier (SA) developed a relatively simple classification model (SA) based on association rules. A simple associative classifier was presented by selecting a resealable number of class association rules for each class. The algorithm finds all the rules in the dataset and sorts them based on support and confidence measures. Then, the strong CARs are grouped according to class label, and finally, the user-specified (intended number of CARs) number of CARs for each class is extracted to build a simple associative classifier.

In J&B method, a thorough search of the entire example space yielded a descriptive and accurate classifier (J&B). To be more specific, CARs are first produced using the APRIORI method. The strong class association rules are then chosen based on how well they contribute to enhancing the overall coverage of the learning set. In the rule selection process, J&B has a halting condition based on the coverage of the training dataset. If it satisfies the user-defined threshold (intended dataset's coverage), it stops the rule-selection process and forms the final classifier.

The algorithm described here extends the researchers' previous work [27][28]. In [27], the CMAC algorithm was introduced; it first generates the CARs by employing the APRIORI algorithm; secondly, the algorithm uses a direct distance metric in the clustering phase of CARs; finally, the cluster centroid approach is applied to select the representative CAR for each cluster, while in [28] CMAC is compared to two similar algorithms, one (DDC) using the same direct distance metric for clustering and covering approach in the representative CAR selection phase; the other algorithm (CDC) using combined (direct and indirect) distance metric with the same covering approach to select the representative CAR. This research presents a similar approach using a combined distance metric (three different metrics are proposed by considering the contribution of direct and indirect measures) in the CAR clustering phase after the CARs are found by using the APRIORI algorithm, and the cluster centroid approach is used to select the representative CAR for each cluster.

Plasse et al. [29] discovered hidden regularities between binary attributes in large datasets. The authors used similar techniques as in the research here: clustering and association rule mining to reduce the number of Ars produced, but the proposed algorithm was totally different. Since there were 3000 attributes in the dataset, their main goal was to cluster (by using the hierarchical clustering algorithm) the attributes to reveal interesting relations between binary attributes and to further reduce the future space. Using the APRIORI algorithm, strong meaningful ARs were generated in the clustered dataset, which can be used for further classification purposes.

In [30], the authors developed a new algorithm based on strong class association rules, which obtained 100% confidence. They directly produced CARs with higher confidence to build a compact and accurate model. A vertical data format [31] was utilized to generate the rule items associated with their intersection IDs. The support and confidence values of the CARs were computed based on the intersection technique. Once the potential CAR is

found for the classifier, the associated transaction will be discarded by using a set difference to eliminate generating redundant CARs. This is a nice related work that differs from the researchers' method in the rule selection stage. More precisely, any clustering technique was used in the rule extraction phase of the proposed model.

The distance-based clustering approach [32] aims to cluster the association rules generated from numeric attributes. They followed the same process to cluster and select the representative rule for each cluster as in the researchers' algorithm. The steps are similar, but the methods used in each step are different. (1) They used the “APRIORI” algorithm to find the association rules from the dataset with numeric attributes; (2) since they are working with numeric attributes, the Euclidean distance metric is used to find similarities between association rules; (3) a representative rule is selected based on coverage, which measures the degree of a certain rule to cover all others.

In [33], researchers proposed a new similarity measure based on the association rule for clustering gene data. They first introduced a feature extraction approach based on statistical impurity measures, such as the Gini Index and Max Minority, and they selected the top 100–400 genes based on that approach. Associative dependencies between genes are then analyzed, and weights to the genes are assigned according to their frequency of occurrences in the rules. Finally, a weighted Jaccard and vector cosine similarity functions are presented to compute the similarity between the generated rules, and the produced similarity measures are applied later to cluster the rules by utilizing the hierarchical clustering algorithm. In this approach, some steps are similar to the researchers' method, but different techniques are used in those steps.

In [34], researchers proposed a novel distance metric to measure the similarity of association rules. The main goal of the research was to mine clusters with association rules. They first generated the association rules by using the “APRIORI” algorithm, one of the most-used algorithms. They introduced an “Hamming” distance function (based on coverage probabilities) to cluster (a hierarchical clustering algorithm is used) the rules. The key difference between the researchers' method and the proposed method is that this research aimed to produce a compact and accurate associative classifier, while its main goal was to measure the quality of the clustering.

In [35], the authors focused on detecting unexpected association rules from transactional datasets. They proposed a method for generating unexpected patterns based on beliefs automatically derived from the data. They clustered the association rules by integrating a density-based clustering algorithm. Features are represented as vectors captured by semantic and lexical relationships between patterns. The clustering phase considers such logical relationships as similarities or distances between association rules. The idea is slightly similar to ours, but used a different clustering technique and cluster association rules, not class association rules.

References

1. Lent, B.; Swami, A.; Widom, J. Clustering association rules. In Proceedings of the Thirteenth International Conference on Data Engineering, Birmingham, UK, 7–11 April 1997; Gray, A., Larson, P., Eds.; pp. 220–231.
2. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. In Proceedings of the VLDB '94 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, 12–15 September 1994; pp. 487–499.
3. Hall, M.; Frank, E. Combining Naive Bayes and Decision Tables. In Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, Coconut Grove, FL, USA, 15–17 May 2008; Wilson, D.L., Chad, H., Eds.; pp. 318–319.
4. Kohavi, R. The Power of Decision Tables. In Proceedings of the 8th European Conference on Machine Learning, Crete, Greece, 25–27 April 1995; pp. 174–189.
5. Hühn, J.; Hüllermeier, E. FURIA: An algorithm for unordered fuzzy rule induction. Data Min. Knowl. Discov. 2019, 19, 293–319.
6. Frank, E.; Witten, I. Generating Accurate Rule Sets Without Global Optimization. In Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA, USA, 24–27 July 1998; Shavlik, J.W., Ed.; pp. 144–151.
7. Quinlan, J. C4.5: Programs for Machine Learning. Mach. Learn. 1993, 16, 235–240.
8. Richards, D. Ripple down Rules: A Technique for Acquiring Knowledge. Decision-Making Support Systems: Achievements, Trends and Challenges for the New Decade; IGI Global: Hershey, PA, USA, 2002; pp. 207–226.
9. Liu, B.; Hsu, W.; Ma, Y. Integrating classification and association rule mining. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 27–31 1998; Agrawal, R., Stolorz, P., Eds.; pp. 80–86.
10. Hu, L.Y.; Hu, Y.H.; Tsai, C.F.; Wang, J.S.; Huang, M.W. Building an associative classifier with multiple minimum supports. Springer Plus 2016, 5, 528.
11. Deng, H.; Runger, G.; Tuv, E.; Bannister, W. CBC: An associative classifier with a small number of rules. Decis. Support Syst. 2014, 50, 163–170.
12. Ramesh, R.; Saravanan, V.; Manikandan, R. An Optimized Associative Classifier for Incremental Data Based On Non-Trivial Data Insertion. Int. J. Innov. Technol. Explor. Eng. 2019, 8, 4721–4726.
13. Thabtah, F.A.; Cowling, P.; Peng, Y. MCAR: Multi-class classification based on association rule. In Proceedings of the 3rd ACS/IEEE International Conference on Computer Systems and Applications, Cairo, Egypt, 6 January 2005; pp. 127–133.

14. Thabtah, F.A.; Cowling, P.; Peng, Y. MMAC: A new multi-class, multi-label associative classification approach. In Proceedings of the Fourth IEEE International Conference on Data Mining, Brighton, UK, 1–4 November 2004; pp. 217–224.
15. Abdellatif, S.; Ben Hassine, M.A.; Ben Yahia, S.; Bouzeghoub, A. ARCID: A New Approach to Deal with Imbalanced Datasets Classification. In Proceedings of the 44th International Conference on Current Trends in Theory and Practice of Computer Science, Lecture Notes in Computer Science, Krems, Austria, 29 February 2018; Volume 10706, pp. 569–580.
16. Chen, G.; Liu, H.; Yu, L.; Wei, Q.; Zhang, X. A new approach to classification based on association rule mining. Decis. Support Syst. 2006, 42, 674–689.
17. Mattiev, J.; Kavsek, B. Coverage-Based Classification Using Association Rule Mining. Appl. Sci. 2020, 10, 7013.
18. Kaufman, L.; Rousseeuw, P.J. Finding Groups in Data: An Introduction to Cluster Analysis; John Wiley and Sons: Hoboken, NJ, USA, 1990.
19. Zait, M.; Messatfa, H. A Comparative Study of Clustering Methods. Future Gener. Comput. Syst. 1997, 13, 149–159.
20. Arabie, P.; Hubert, L.J. An Overview of Combinatorial Data Analysis. In Clustering and Classification; Arabie, P., Hubert, L.J., Soete, G.D., Eds.; World Scientific Publishing: Singapore, 1996; pp. 5–63.
21. Dahbi, A.; Mouhir, M.; Balouki, Y.; Gadi, T. Classification of association rules based on K-means algorithm. In Proceedings of the 4th IEEE International Colloquium on Information Science and Technology, Tangier, Morocco, 24–26 October 2016; Mohajir, M.E., Chahhou, M., Achhab, M.A., Mohajir, B.E., Eds.; pp. 300–305.
22. Ng, T.R.; Han, J. Efficient and Effective Clustering Methods for Spatial Data Mining. In Proceedings of the 20th Conference on Very Large Data Bases (VLDB), Santiago, Chile, 12 September 1994; Kaufmann, M., Ed.; pp. 144–155.
23. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In Proceedings of the ACM-SIGMOD International Conference on Management of Data, Montreal, Canada, 1 June 1996; pp. 103–114.
24. Theodoridis, S.; Koutroumbas, K. Hierarchical Algorithms. Pattern Recognit. 2009, 4, 653–700.
25. Liu, B.; Chen, G. The Association Rules Algorithm Based on Clustering in Mining Research in Corn Yield. In Computer and Computing Technologies in Agriculture X. CCTA 2016. IFIP Advances in Information and Communication Technology; Li, D., Ed.; Springer: Dongying, China, 2019; Volume 509, pp. 268–278.

26. Sunita, D.A.; Lobo, J. Combination of Clustering, Classification & Association Rule based Approach for Course Recommender System in E-learning. Int. J. Comput. Appl. 2012, 39, 8–15.
27. Mattiev, J.; Kavšek, B. CMAC: Clustering Class Association Rules to Form a Compact and Meaningful Associative Classifier. In Machine Learning, Optimization, and Data Science. LOD-2020. Lecture Notes in Computer Science; Nicosia, G., Ed.; Springer: Siena, Italy, 2020; Volume 12565, pp. 372–384.
28. Mattiev, J.; Kavšek, B. Distance-based clustering of class association rules to build a compact, accurate and descriptive classifier. Comput. Sci. Inf. Syst. 2021, 18, 791–811.
29. Plasse, M.; Niang, N.; Saporta, G.; Villeminot, A.; Leblond, L. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput. Stat. Data Anal. 2007, 52, 596–613.
30. Thanajiranthorn, C.; Songram, P. Efficient Rule Generation for Associative Classification. Algorithms 2020, 13, 299.
31. Ogihara, Z.P.; Zaki, M.; Parthasarathy, S.; Ogihara, M.; Li, W. New algorithms for fast discovery of association rules. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA, USA, 14–17 August 1997.
32. Du, X.; Suzuki, S.; Ishii, N. A Distance-Based Clustering and Selection of Association Rules on Numeric Attributes. In New Directions in Rough Sets, Data Mining, and Granular-Soft Computing; RSFDGrC 1999. Lecture Notes in Computer Science; Zhong, N., Skowron, A., Ohsuga, S., Eds.; Springer: Berlin/Heidelberg, Germany, 1999; Volume 1711, pp. 423–432.
33. Sethi, P.; Alagiriswamy, S. Association Rule Based Similarity Measures for the Clustering of Gene Expression Data. Open Med. Inform. J. 2010, 4, 63–73.
34. Kosters, W.; Marchiori, E.; Oerlemans, A. Mining Clusters with Association Rules. In Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 1999.
35. Bui-Thi, D.; Meysman, P.; Laukens, K. Clustering association rules to build beliefs and discover unexpected patterns. Appl. Intell. 2020, 50, 1943–1954.