Abstract

Associative Classification leverages Association Rule Mining (ARM) to train Rule-based classifiers. The classifiers are built on high quality Association Rules mined from the given dataset. Associative Classifiers are very accurate because Association Rules encapsulate all the dominant and statistically significant relationships between items in the dataset. They are also very robust as noise in the form of insignificant and low-frequency itemsets are eliminated during the mining and training stages. Moreover, the rules are easy-to-comprehend, thus making the classifier transparent.

Conventional Associative Classification and Association Rule Mining (ARM) algorithms are inherently designed to work only with binary attributes, and expect any quantitative attributes to be converted to binary ones using ranges, like “Age = [25, 60]”. In order to mitigate this constraint, Fuzzy logic is used to convert quantitative attributes to fuzzy binary attributes, like “Age = Middle-aged”, so as to eliminate any loss of information arising due to sharp partitioning, especially at partition boundaries, and then generate Fuzzy Association Rules using an appropriate Fuzzy ARM algorithm. These Fuzzy Association Rules can then be used to train a Fuzzy Associative Classifier.

To mine Fuzzy Association Rules, we need a robust, fast, and scalable Fuzzy ARM algorithm. The only fuzzy ARM algorithms which are most popular, as of today, are various fuzzy adaptations of the original crisp Apriori algorithm. Fuzzy Apriori, like original Apriori, is a slow algorithm, especially for most medium-sized (500K–1M), and large (1M–10M) datasets. To this end, we present a new fuzzy ARM algorithm called FAR-Miner, which is loosely based on the Oracle and ARMOR algorithms. It is designed for very fast and efficient performance on any dataset, especially medium-sized to large ones. From our experiments on a medium-sized real-life dataset and a large real-life dataset, we show that FAR-Miner is 8–19 times faster for large datasets, and 6–10 times for medium-sized datasets as compared to fuzzy Apriori.

Fuzzy association rules can also be derived from high-dimensional numerical datasets, like image datasets, in order to train fuzzy associative classifiers. But, because of the peculiarity of such datasets, traditional fuzzy ARM algorithms are not able to mine rules from them efficiently, since such algorithms are meant to deal with datasets with relatively much less number of attributes/dimensions. Hence, we propose FAR-HD which is a fuzzy ARM algorithm designed specifically for very large high-dimensional datasets. FAR-HD processes fuzzy frequent itemsets in a tree-like (depth first search – DFS) manner using a two-phased multiple-partition tidlist-based strategy. Additionally, FAR-HD uses Fuzzy Clustering to convert each numerical vector of the original input dataset to a fuzzy-cluster-based representation, which is ultimately used for the actual fuzzy ARM process. The performance of FAR-HD far exceeds that of Fuzzy Apriori and FAR-Miner on various standard public image datasets (varying in size from
medium to very large), based on metrics like execution time, maximum memory used, and number of page faults generated.

But, before any fuzzy ARM algorithm can be used, the original dataset (with crisp attributes) needs to be transformed into a form with fuzzy attributes. This pre-processing is done through our algorithm called FPrep, which involves using fuzzy clustering to generate fuzzy partitions. Then, these partitions are used to get a fuzzy version (with fuzzy records) of the original dataset. The fuzzy data is used to mine fuzzy association rules using a suitable algorithm like FAR-Miner. We also show that FPrep is much faster than other such comparable transformation techniques, which in turn depend on non-fuzzy techniques, like hard clustering (CLARANS and CURE). Moreover, we illustrate the quality of the fuzzy partitions generated using FPrep, and the number of frequent itemsets generated by a fuzzy ARM algorithm when preceded by FPrep.

Association Rules (fuzzy or crisp) mined from various datasets and domains can be used to build highly accurate, effective, and robust Associative Classifiers. We describe three Associative Classifiers (one crisp – SEAC and two fuzzy – FSEAC and FACISME) in this thesis. We also discuss the application of our Associative Classification approach in two specialized domains (Identification of object classes in images and Look-alike Modeling for online ad-targeting), and how customized classifiers are built for these domains.

SEAC is a holistic, straightforward, and fast classification algorithm, which uses association rules directly for classification, after two phases of processing and pruning. The first phase performs association rule mining (ARM) and subsequent classifier building globally by taking the whole dataset into account. The second phase does the same albeit locally, i.e. only for those classes which are either under-represented or not represented at all in the first phase. Doing so bolsters the accuracy, and drastically reduces the chances of an unclassified tuple, with a rare consequent class label, being misclassified or not classified at all. SEAC uses very few parameters, which can be very easily and effortlessly configured. It relies on information gain and entropy to build a set of optimal rules. Thus, its accuracy is very good as compared to other popular contemporary state-of-the-art classification algorithms, associative as well as non-associative ones.

FSEAC which extends the crisp Associative Classifier SEAC, while preserving its advantages, such as simplicity and ease-of-use, along with comprehensible rule sets. It can be used to train on datasets with any type of attributes binary or numerical. In FSEAC numerical attributes would be replaced by fuzzy ones. Doing so ensures that there is no loss of information whatever the value of any numerical attribute. Thus, the inherent uncertainty that is present in numerical data would also appropriately be taken care of. Like SEAC, FSEAC is designed to function in a two-phased manner so as to achieve maximum accuracy and maximum representation of all possible class labels involved in the domain. Largely because of the fuzzy flavour, FSEAC provides better accuracy as compared to SEAC, as is evident from our experimental studies. This is because of the ability of FSEAC in dealing with numerical attributes in an effective manner through soft partitions. Our experimental studies show that FSEAC, when compared with 16 state-of-the-art classification algorithms on 23 disparate UCI-ML datasets, provides best accuracy for 17 datasets (absolute difference with the second best algorithm of 0.5% for 12 of these 17 datasets), and is in the top three for three other datasets.
Next, we have also developed a Fuzzy Associative Classification algorithm called FACISME. It is based on maximum entropy, and uses iterative scaling, both of which lend a very strong theoretical foundation to the algorithm. Entropy is one of the best measures of information, and maximum-entropy-based algorithms do not assume independence of parameters in the classification process. Thus, FACISME provides very good accuracy, and can work with all types of datasets, irrespective of the type of attributes (numerical or binary) involved.

The techniques used in SEAC and FSEAC form the nucleus of our Associative Classification approach, both crisp and fuzzy. Last, we show applications of our approach in two domains by creating customized classifiers based on this approach. The application is to identify the presence of object classes in image datasets (containing only numerical features), and the other domain is in the Look-alike Modeling for User Targeting in Display Advertisements.

For the first domain, we present I-FAC, a novel Fuzzy Associative Classification algorithm for object class detection in images using interest points. In object class detection, the negative class \( C_N \) is generally vague \( (C_N = U - C_P) \); where \( U \) and \( C_P \) are the universal and positive classes respectively. But, image classification necessarily requires both positive and negative classes for training. I-FAC is a single-class image classifier that relies only on the positive class for training. Because of its fuzzy nature, I-FAC also handles polysemy and synonymy (common problems in most crisp (non-fuzzy) image classifiers) very well. The performance of I-FAC, as adjudged from its ROC curve is very good, especially at lower false-positive-rates when its recall is even better.

We also present an associative classification algorithm for visual concept detection using interest points (color descriptors). This algorithm is an extension of I-FAC, and relies on fuzzy (soft) partitions to generate a per-descriptor record model which is used for training associative classifiers. We use the first 10000 images of the MIR dataset (with 38 visual concepts – 14 regular annotation topics and 24 pre-annotations) for evaluating our algorithm. As associative classification leverages frequent patterns mined from a given dataset, its performance as adjudged from average area under curve (AUC) values over three-fold cross validation is very good, especially for regular annotation topics and for “scenic” concepts. From an empirical perspective (on the MIR dataset), the performance of our algorithm is much better, than that of either bag-of-words (Linear SVM and SVM RBF Kernel) or Linear SVM based on fuzzy (soft) partitions – both benchmark approaches are based on interest points (color descriptors).

Online advertising world is constantly becoming more competitive, with advertisers now wanting not only to optimize for Click-through-rate (CTR), but also for conversions, which is their ultimate goal. Given a seed list of users who have previously converted on an ad-campaign, the aim of Look-alike Modeling is to find a new set of look-alike users who have a high probability of converting, and thus can be targeted. Advertisers want to optimize for conversions which typically occur at a rate in the order of one conversion for every 1M impressions. Custom campaign-based Look-alike models, that we present can be used to target users with conversion as the goal, not just clicks. Training custom Look-alike models for advertisers is a challenging problem, especially for tail campaigns. Campaigns with very few conversions (positive class) are called as tail campaign. We describe a Hadoop-based Associative Classification approach to Look-alike Modeling for tail campaigns. Pairs of features are used to derive rules to build a Rule-based Associative Classifier, with the rules being sorted by frequency-weighted
log-likelihood ratio (F-LLR). The top \( k \) rules, sorted by F-LLR, are then be applied to any test record to score it.

Thus, in this thesis we have described a broad framework for Associative Classification, especially Fuzzy Associative Classification, which starts from the pre-processing (in case of fuzzy classification) of the original dataset to an appropriate fuzzy format for ARM. Then, association rules are mined using our ARM algorithms. These association rules are then leveraged to build an Associative Classifier (fuzzy or crisp). This is done using our associative classification algorithms. Last, this approach can be applied to build associative classifiers in specialized domains. This can be done using customized associative classification algorithms, which are based on our associative classification approach and framework.