Chapter 2

Overview of Our Approach

In this section, we describe the details of the design and implementation of our approach. The whole process is divided into two broad phases – first phase in which fuzzy ARM, with appropriate pre-processing, is done, and the second phase in which the actual associative classifier training is done. A high-level overview of the steps followed in our associative classification approach are:

- Pre-processing to generate fuzzy dataset (for fuzzy associative classifiers) using FPrep
- Classification Association Rules (CARs) mining using FAR-Miner or FAR-HD
- CARs pruning and classifier training using SEAC or FSEAC

2.1 Fuzzy Pre-Processing and Pattern Mining using ARM

The first phase is meant exclusively to mine frequent patterns, which encapsulate all trends pertaining to the problem at hand. Thus, any task, like classification, depending on such frequent patterns would be done in a comprehensive manner, without anything important being missed. These frequent patterns are represented in the form of association rules. Besides using conventional crisp association rules, we use fuzzy association rules as well. By coupling fuzzy logic and association rules, we are able to deal with all types of attributes present in any given dataset without any loss of information. Previously, with crisp association rules, we would have had to use sharp partitioning of numerical attributes in order to get binary attributes derived from such numerical attributes. But by doing so, we introduce loss of information. For example, given a particular partition “Advertisement cost = [Rs. 100, Rs. 400]”, both Rs. 150 and Rs. 399 would fit in the partition, though both these values are far apart.

Using fuzzy logic, we create fuzzy partitions and determine the membership ($\mu$) of each numerical data point in these fuzzy partitions. Thus, each numerical point would belong partially to more than one partition depending on its actual value. For example, given two fuzzy partitions “Advertisement cost = Low” and “Advertisement cost = Medium”, Rs. 150 would belong to both these partitions with $\mu_{\text{low}} = 0.7$ and $\mu_{\text{medium}} = 0.2$, where as Rs. 399 would have $\mu_{\text{low}} = 0.05$ and $\mu_{\text{medium}} = 0.9$. Thus, we see that fuzzy attributes provide a more correct reflection of the original numerical attributes as compared to binary attributes derived by sharp partitioning. These fuzzy attributes are clubbed with any binary attribute in the original dataset. Using this new set of attributes we convert each transaction of
the original dataset to get a transformed fuzzy dataset, which is then used subsequently to determine the actual fuzzy association rules. These fuzzy association rules better represent the frequent patterns, present in the original dataset, as compared to just conventional association rules. This whole process of creating fuzzy partitions, and then using them to obtain a fuzzy version of the original dataset forms a part of the fuzzy pre-processing done in the first phase. The actual pattern mining process ensues after this pre-processing.

The pre-processing for high-dimensional datasets, like image datasets, is slightly different and a bit more straightforward. Fuzzy clustering (like Fuzzy c-Means – FCM) is applied on each record or feature vector of the training dataset. The fuzzy clusters thus obtained are used to transform each feature vector into a corresponding record containing cluster-ids and its memberships ($\mu$) in the clusters. The transformed fuzzy records are then mined for frequent itemsets and fuzzy association rules.

Therefore, in the first phase we mine (FAR-Miner algorithm for general datasets and FAR-HD for high-dimensional datasets) for frequent patterns, which are ultimately represented in the form of association rules. These frequent patterns and association rules form the driving force for the ensuing classification process in the second phase.

### 2.2 Associative Classification and Association Rules

The second phase leverages the frequent patterns and association rules generated previously in the first phase. Associative classification can be done using the algorithms we have developed, depending on the type of dataset and domain.

We use the actual association rules also for the final classification process (using SEAC, our associative classifier). The two major hurdles in using these association rules directly, namely their huge volume and redundancy are dealt with appropriately by SEAC. It uses a constrained exhaustive mining approach to obtain the best possible results in terms of accuracy, and does not use any greedy approach. SEAC deals with redundant association rules through an effective and simple pruning technique which also helps in cutting down the number of rules which finally form a part of the classifier. The classifier is built in a two-phased manner so as to achieve maximum accuracy and maximum representation of all possible class labels involved in the domain.

For general datasets (having a mix of binary and numerical features), we use fuzzy association rules for classifier training. For this we use FSEAC which extends the crisp Associative Classifier SEAC, while preserving its advantages, such as simplicity and ease-of-use. FSEAC can be used to build an associative classifier based on the association rules generated by FAR-Miner.

We also have customizations of our associative classification approach for specialized domains. For example, for high dimensional data, like images, we have another fuzzy associative classification algorithm called I-FAC. I-FAC and another algorithm which is its extension, can be used to build classifiers using fuzzy association rules mined (using FAR-HD) from high dimensional vectors (like feature vector representation of images). We also present another algorithm which is an extension of I-FAC, and which adapts associative classification for finding various visual concepts (like car, face, tree, and beach) in images, by leveraging Color Descriptors that can be extracted from images. Another customization is the
application of our associative classification approach in Look-alike Modeling for Online User-Targeting of Ads.

Thus, our ARM and associative classification require minimal user-intervention while training or actual classification, apart from being highly accurate and requiring very little time for training (in the order of few minutes) even for huge datasets. Moreover, all our classifiers require very little time (in the order of few milliseconds) for the actual classification process. As more classification-based applications today are online, the classification time assumes a lot of importance, and having very low classification time would be of great advantage.