Fuzzy Meta-Learning: Preliminary Results*

Grigorios Tsoumakas
Department of Informatics
Aristotle University of Thessaloniki
54006 Thessaloniki, Greece
greg@csd.auth.gr

Ioannis Vlahavas
Department of Informatics
Aristotle University of Thessaloniki

Abstract

Learning from distributed data is becoming in our times a necessity, but it is also a complex and challenging task. Approaches developed so far have not dealt with the uncertainty, imprecision and vagueness involved in distributed learning. Meta-Learning, a successful approach for distributed data mining, is in this paper extended to handle the imprecision and uncertainty of the local models and the vagueness that characterizes the meta-learning process. The proposed approach, Fuzzy Meta-Learning uses a fuzzy inductive algorithm to meta-learn a global model from the degrees of certainty of the output of local classifiers. This way more accurate models of collective knowledge can be acquired from data with application both to inherently distributed databases and parts of a very large database. Preliminary results are promising and encourage further research towards this direction.

1 Introduction

Nowadays, information systems are getting larger and richer in content (environmental information systems, multimedia databases, satellite data, the world wide web) and at the same time they also grow in number and diversity. These systems are usually dynamic, in the sense that information is added, removed and altered over time. In certain domains, knowledge itself changes over time. This phenomenon known as concept drift appears often in medical, environmental, financial and other information systems that are concerned with complex and evolving physical systems. Furthermore, advances in data communication have allowed the interconnection of modern information systems through the Internet or private networks. Some systems are composed of physically distributed smaller units and use networks for communication and cooperation.

Discovering interesting patterns from data residing in systems with the aforementioned properties is a challenging task. One has to deal effectively with the size, distribution, complexity and evolution of data. In recent years, several approaches that deal with the distributed nature of the learning problem have been developed from the relatively new field of Distributed Data Mining (Kargupta & Chan, 2000). These include both new algorithms that allow for distributed learning, as well as new system architectures that support the actual learning process. One of the most promising lines of research in this field is *Meta-Learning* (Chan & Stolfo, 1993), which is a methodology for deriving a single global classification model by learning from multiple local classifiers. A problem with Meta-Learning and other approaches of this field is that they do not take into account the uncertainty and ambiguity involved in the distributed learning environment.

In this paper, Meta-Learning is extended to handle the imprecision and uncertainty of the local models and the vagueness that characterizes the meta-learning process. The proposed approach, *Fuzzy Meta-Learning*, uses a fuzzy inductive algorithm to meta-learn a global model from the degrees of certainty of the output of local classifiers. This way more accurate models can be acquired from data with application both to inherently distributed databases and parts of a very large database.

The rest of this paper is organized as follows. Section 2, outlines related work on meta-learning as an approach to learning from distributed data. Section 3, introduces Fuzzy Meta-Learning and Section 4 exhibits experimental results. Finally, Section 5 recapitulates this work and points to future research directions.

* This work is partially funded by the Greek Secretariat for Research and Technology under project 9513514, PDE, EPET II, Action 2.5.

2 Related Work

There has been a lot of research activity recently on learning from distributed data. The main concept of this field is essentially learning a variety of models from disjoint data sets using the same or different algorithms and then combining them to discover global consistent knowledge. Although distributed learning is by definition applied to inherently distributed databases, it also provides the means to scale-up learning to very large databases. This requires different architecture in order to manually distribute the data to several processors but at the same times allows for more communication between the learning processes, as they will probably be running on computers of the same network or processors of the same machine. Distributed learning can also lead to an increase in accuracy, especially when different learning algorithms with different biases are used. This is due to the fact that each algorithm may compensate the inefficiencies of the other.

A very attractive method for distributed learning is Meta-Learning (Chan & Stolfo, 1993). This approach is based on combining the output of several classification systems (base classifiers) into a final classifier. The combination involves *learning* the actual way that the output of several classifiers correlates with the true class given a training example. There are two approaches to this idea, the *arbiter* and the *combiner*. In the first approach, apart from the base classifiers, there is another one called the arbiter. This classifier decides the final output when the base classifiers cannot reach a consensus. In the combiner approach, a single classifier outputs the result based on the results of the base classifiers. In both approaches, the arbiter or combiner prediction is learned using as training data the output of the base classifiers.

One of the important advantages of this method is its independence of the actual algorithm used for base classifier learning. Furthermore, it can be effectively used both in a distributed environment, and for scaling up learning to very large databases. It is also combined with an excellent agent-based architecture, which offers the ability to deal with heterogeneous environments (Prodromidis et al, 1999). However, this method has the disadvantage that the meta-level model is not comprehensible. It describes the way base classifiers correlate with the correct class and not the way data correlate with the correct class. In addition, the uncertainty that underlies the base model predictions is not taken into account.

An extension to this work is *Knowledge Probing* (Guo & Sutiwaraphun, 1999). This method builds on the idea of meta-learning and in addition uses an independent data set, called the probing set, in order to discover knowledge about the data. The output of a meta-learning system on this independent data set together with the attribute value vector of the same data set are used as training examples for a learning algorithm that outputs a final model. In this elegant way, the disadvantage of having a black box is overcome and the result is a transparent predictive model. However, the choice of size and origin of the probing set are issues that have to be thoroughly investigated, especially in the context of a distributed environment. As in the previous approach, the uncertainty of learning from the predictions of the base classifiers is not considered.

The theory of Fuzzy Sets (Zadeh, 1965) has in the past been successfully applied to learning from data. It has been used for a range of learning tasks including discovering classification knowledge, association rules and clustering (Pedrycz, 1997; Berthold, 1999). Amongst the main strengths of Fuzzy Logic for knowledge discovery is the ability to model linguistic, vague or imprecise information. For this reason, we believe that it can have a prominent role in learning from distributed information systems, a field where it has not been systematically used yet.

3 Fuzzy Meta-Learning

The approach of Fuzzy Meta-Learning (FML) is motivated by the fact that learning collective knowledge is intuitively a vague process. None of the local models can be absolutely perfect in real-world problems and in addition might include knowledge that is not globally consistent, but only describes local properties of the data. For this reason, FML proposes a) an explicit description of this uncertainty in the output of local models and b) the use of a fuzzy inductive algorithm to learn a global model from the local models.

3.1 The Methodology

The framework of FML can be broken down to two main phases, based on the paradigm of classical Meta-Learning:

1. Use one or more (fuzzy or crisp) machine learning algorithms to produce an ensemble of base predictive models over the distributed data sets (or parts of a very large database).

2. Meta-learn a model that describes the behavior of all the base models, using a *fuzzy* inductive algorithm on a separate evaluation data set.

In the first phase, a learning algorithm is used locally at each distributed site, resulting in an ensemble of local models. It is possible to have different algorithms at different sites, as long as they produce classifiers that associate a degree of certainty along with the resulting class. The learning algorithms can either be fuzzy or crisp depending on the domain characteristics and data properties. Note that even crisp classifiers can associate a degree of certainty to the resulting class (decision trees, neural networks) and can output more than one class (pruned rules of decision trees, neural networks). Fuzzy classifiers output more than one class with a varying degree of certainty by nature. Therefore, they are expected to result in better performance of the whole methodology than crisp ones.

In the second phase, the base models are evaluated on a common independent evaluation data set. For each example of this data set the output of the base models is combined in order to create a meta-level training example based on the following proposed *meta-fuzzy* scheme: The degrees of certainty of the classifications of all the base models with respect to all classes form the attribute vector of the meta-level training example. The correct class of the examples follows. If a classifier outputs more than one degree of certainty for the same class then the maximum degree is retained.

Example	Attribute vector	Class
x	attrvec(x)	class(x)
x1	$attrvec_{x1}$	good
x2	$attrvec_{x2}$	bad
x3	$attrvec_{x3}$	good

Sample evaluation set

Classifier predictions	
$C_1(x)$	$C_2(x)$
good	good
good	bad
bad	bad

Classifier predictions

Attribute Vector	Class
attrvec(y)	class(y)
good, good	good
good, bad	bad
bad, bad	good

Meta-level training set

Classifier predictions			
$C_1(x)$		$C_2(x)$	
good	Bad	good	bad
0.7	0.2	0.8	0.3
0.6	0.9	0.6	0.8
0.4	0.6	0.4	0.6

Certainty of classifier predictions

Attribute Vector	Class
attrvec(y)	class(y)
0.7, 0.2, 0.8, 0.3	good
0.6, 0.9, 0.6, 0.8	bad
0.4, 0.6, 0.4, 0.6	good

Meta-level training set

Figure 1: Structuring Schemes for Meta-Learning

An example of the process of creating meta-level training examples using the proposed scheme as well as using the classical meta-learning scheme is depicted in Figure 1. There is one class to be predicted with two possible values, good and bad. There are two learned base classifiers $C_1(x)$ and $C_2(x)$, whose predictions are going to be combined for structuring the meta-level training examples. In the first table there is an evaluation data set with 3 instances. The following two tables show the meta-learning scheme called *meta-class*, which structures the meta-level examples in the right table with the dominating class predictions of the two base classifiers seen in the left table. The proposed *meta-fuzzy* scheme is depicted at the bottom of the figure. The meta-level training examples in the right table are structured in this case with the degrees of certainty of the base classifiers predictions with respect to all classes seen in the left table. This scheme intuitively allows for finer learning the base classifier correlation with respect to the true class, since information about the base classifier output is more and of higher quality.

Once the meta-level training examples are constructed, the next step in the second phase is the induction of the final classifier. FML introduces here the use of a fuzzy inductive algorithm for learning the global predictive model directly from the meta-level training examples. The rationale behind this approach is that the problem of learning collective knowledge is a complex task characterized by ambiguity and uncertainty. These properties come from:

- The intricacy of the original learning problem itself. Real-world learning problems have their own complexity, which is propagated to the meta-level learning problem. It is usually very difficult to induce a perfect classifier from the original data, and thus it is common to settle with a classifier of an acceptable performance. This means that the training data used for meta-learning will be uncertain, as output of imperfect base classifiers.

- The locality of the base models. In the inherently distributed databases scenario the classifiers are trained on the local data of each database and therefore might include knowledge that has local characteristics and is not globally consistent. In the very large database scenario, the learning process is restricted by the fact that it examines a sample of the data. In both cases, uncertainty is introduced in the meta-level training examples.

An issue at this stage is what kind of fuzzy learning algorithm to use. First of all, there is no background domain knowledge for the meta-learning task. Even if such knowledge existed, it would be different for each domain and classifier used. In addition, the comprehensibility of the final model is not important, because it does not provide knowledge about the data, rather describes the correlation of the base classifiers with respect to the true class. Therefore an adaptive fuzzy learning algorithm would be more effective for FML than an algorithm that uses a fixed set of fuzzy sets determined prior to the learning process. Adaptive algorithms use techniques from the field of Computational Intelligence and Soft Computing like genetic algorithms and neural networks to fine-tune the fuzzy sets used for the granulation of the attribute space. In this way they can increase the predictive performance of the induced fuzzy model.

After the two phases of FML described above, the global model is available for classification of examples with unknown class values. A new example has first to be classified by all the base models. Then their output is used as input to the global predictive model, which in turn outputs the final classification. The extra cost of passing new examples through all base classifiers is negligible, because all base models are made locally available after the end of the first phase. The broadcast of all models to all local sites comes without much communication overhead, as models are small in size.

3.2 Scaling up to large classes and distributions

A problem that arises when using classifier results with respect to all classes in order to structure the meta-training examples in domains with large number of classes, is that the attribute-vector length grows greatly as the number of base classifiers is increased. Consider for example the domain of English letter recognition, where there are 26 possible classes (A…Z). If training data are split in 4 different sets, then the meta-level training data set will have 104 attributes. This is not a small number of attributes for a class of learning algorithms. If we also take into account that fuzzy algorithms have usually the disadvantage of great complexity with respect to the length of attributes, we can understand that FML with the proposed scheme in such domains becomes infeasible.

A workaround for this problem is to meta-learn the aggregation of the classifier predictions over all classes. This way, the length of the attribute vector will always be equal to the number of classes of the domain. Unfortunately, this comes at the expense of losing the fine grain representation of the base classifiers knowledge. A robust method to aggregate classifiers that return measurements along with the result is averaging (Lam, 2000). Therefore, we also propose the following *meta-fuzzy-average* scheme for FML on domains with large number of classes: The averages of the certainty degrees of all base classifiers for each class form the attribute vector of the meta-level training example. The correct class follows as before. If a classifier outputs more than one degree of certainty for the same class then the maximum degree is retained.

4 Experimental Results

The approaches of Meta-Learning and FML were evaluated on two real-world data sets from the Machine Learning Repository at the University of Irvine, California (Blake & Merz, 1998). These were the *adult* and *chess* data sets, large enough (> 1000 examples) to simulate distributed environment. Only two domains were selected at this stage of our research to investigate the performance of the suggested methodology.

The setup of the experiments was the following: Each dataset was randomly split to 25% of evaluation examples and another 75% that was used for distribution to N training sets. In the first phase, each of those sets is used for base model learning using the c4.5rules program (Quinlan, 1993). We used this base learner because it can be applied to categorical, numerical and mixed domains, it is fast and effective and it produces classifications, which sometimes may conflict, with a degree of certainty.

In the second phase, the produced rule sets are used to classify the evaluation examples and the classifications are combined with a suitable scheme to form the meta-level training examples. 75% of these examples are used for training the meta-classifier and the rest for testing it. Meta-learning was performed using the c4.5 program and the meta-class scheme, while FML was performed using the Ruleind program (Shen & Chouchoulas, 1999) with the meta-fuzzy scheme. Ruleind is a non-adaptive grid-based fuzzy learning program based on an algorithm by Lozowski et al. (1993), which demands a pre-granulation of the attribute space. In FML, the attributes are the certainty factors and their domain spans the real number space from 0 to 1. We used the same ad-hoc granulation of the space with three triangular fuzzy sets for all experiments, as seen in Figure 2.

As discussed in the previous section, adaptive fuzzy learning algorithms can have better performance for FML.

Ruleind was used nonetheless, because it was the only available robust implementation of a fuzzy learning algorithm. However, the ad-hoc granulation of the attribute space is expected to negatively affect the predictive performance of Ruleind. Therefore, in order to get a secondary measure of the meta-fuzzy scheme's accuracy impact, c4.5 was also used for meta-learning.

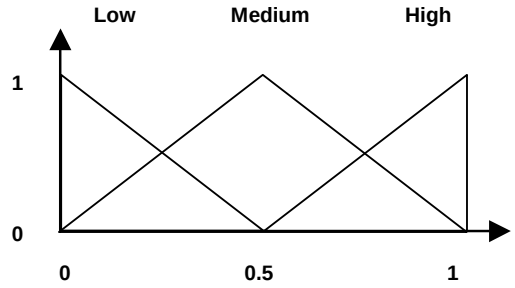

Figure 2: Granulation of the Certainty Degree Space

Table 1, summarizes the results of experiments over 10 validation runs. The second column shows the number of examples while the third the number of classes. The 'S' column shows the number of disjoint data sets that the original data were split into and 'Accuracy' shows the predictive accuracy of the meta-learners with 'C' and 'F' standing for the meta-class and meta-fuzzy combination schemes respectively.

Table 1: Results with the meta-fuzzy Scheme

Data	Inst.	C	S	Accuracy		
				C4.5-C	Ruleind-F	C4.5-F
Chess	3169	2	3	97.82%	98.45%	98.40%
Adult	45000	2	4	86.05%	85.37%	86.35%

As we see from the results in the first dataset, the use of the meta-fuzzy scheme led to higher accuracy than the meta-class scheme. In addition, using the fuzzy learning algorithm had better results even from the state-of-the-art classification system c4.5 with any of the used schemes. In the second dataset, the use of the meta-fuzzy scheme led again to the highest accuracy but using c4.5 and not the fuzzy algorithm. The reason behind this phenomenon is probably that the ad-hoc granulation used along with Ruleind was not effective for this domain. The meta-fuzzy scheme did perform better than meta-class when c4.5 was used, which means that it indeed carries more information, but the fuzzy algorithm could not exploit it. This emphasizes the need for an adaptive algorithm that could use neural networks or genetic algorithms techniques for optimization of the fuzzy sets.

In order to test the meta-fuzzy-average scheme we experimented with two more data sets that have large number of classes. These were the 'Satellite' and 'Segment' data sets again from the Machine Learning Repository. The setup of the experiments was the same as in the case of testing the meta-fuzzy scheme. Table 2 presents the results.

Table 2: Results with the meta-fuzzy-average Scheme

Data	Inst.	C	S	Accuracy		
				C4.5-C	Ruleind-FA	C4.5-FA
Satellite	6435	7	4	84.43%	82.25%	85.32%
Segment	2310	7	3	93.96 %	94.23%	94.03%

Accuracy measurements show that the proposed methodology is more effective. Averaging is performing well for combining measures that classifiers return along with the result, and meta-learning this combination is more informative than meta-learning the dominating class values, at least for domains with large number of classes. However, the use of a fuzzy algorithm for meta-learning showed again performance problems with the first data set.

5 Conclusions and Future Work

In this paper, the Fuzzy Meta-Learning methodology for distributed data mining was introduced and preliminary results from its application to real-world data were presented. As the experimental results exhibited, it leads to models of collective knowledge with increased accuracy. The use of the proposed schemes that take into consideration the uncertainty of the base models has repeatedly led to meta-models with increased predictive performance. The use of a fuzzy inductive algorithm for meta-learning has also shown promising signs in terms of accuracy. Although no general claims can at this point be made about the efficiency of this approach due to limited experiments, the results are encouraging for further research.

5.1 Other advantages of Fuzzy Meta-Learning

Apart from increased accuracy, FML carries intact the advantages of Meta-Learning as an extension of it. First of all, it increases efficiency by learning in parallel from smaller parts of a large database. Therefore, it can be used effectively for scaling up data mining. It can be used in a hierarchical manner with multiple meta-levels of learning, which means that theoretically it can scale up as much as necessary.

In addition, it is independent of the local learning algorithms used, as long as they produce classifiers, and thus allows the use of more than one learning algorithms. In this way it can lead to global models with increased accuracy due to the combination of learning algorithms with different learning biases. However, in order for FML to be fruitful, algorithms that output more than one class with a numerical degree of certainty are preferable.

Finally, it is very suitable for knowledge discovery from inherently distributed databases. Communication and synchronization costs for gathering data in a central server are avoided. In the case of dynamic and evolving systems, models learned in the past can be easily combined with models learned from newer batches of data. Furthermore, Meta-Learning is applicable in the case where the databases concern data that are sensitive and private to competitive organizations. It allows the induction of collective knowledge for common profit of the organizations, without the need for sensitive data to leave their servers. This is especially important for a range of application like fraud detection and e-commerce.

5.2 Future Work

The performance of exhaustive experiments with respect to more domain problems, the use of fuzzy algorithms for base learning, and most importantly the use of an adaptive fuzzy learning algorithm for meta-learning, are included in our immediate research plans. In addition, we plan to perform experiments that investigate the performance of the suggested methodology with respect to the number of distributed sites or parts of a very large database.

A further step would be the implementation of a system for Fuzzy Meta-Learning. Specifically we plan to utilize the new and fast growing architecture of peer-to-peer computing. Peer-to-peer supports the exchange of data in a distributed environment directly from machine to machine either with the use of a central coordination server for addressing or with dynamic discovery based on broadcasting. It has been successfully used by systems that involve analysis of very large data sets (SETI at home, PC-Philanthropy initiative) as well as sharing of terabytes of multimedia data between users (Gnutella, Napster).

References

Berthold, M. (1999). Fuzzy Logic. In Berthold M. & Hand, D. (Eds.), *Intelligent Data Analysis: An Introduction*, pages 269-298. Springer-Verlag.

Blake, C. L., & Merz, C. J. (1998). *UCI Repository of machine learning databases*. Internet resource, available: http://www.ics.uci.edu/~mlearn/MLRepository.html.

Chan, P., & Stolfo, S. (1993). Meta-Learning for multistrategy and parallel learning. *Second International Workshop on Multistrategy Learning*.

Guo, Y., & Sutiwaraphun, J. (1999). Probing Knowledge in Distributed Data Mining. *In Proceedings of PAKDD*, Beijing, China.

Kargupta, H., & Chan, P. (Eds.). (2000). *Advances in Distributed and Parallel Knowledge Discovery*. AAAI/MIT Press.

Lam, L. (2000). Classifier Combinations: Implementations and Theoretical Issues. In Kittler J. & Roli F. (Eds.). *Multiple Classifier Systems*. Lecture Notes in Computer Science series, Springer-Verlag.

Lozowski, A., Cholewo, T. , & Jurada, J. (1996). Crisp rule extraction from perceptron network classifiers. *In Proceedings of the International Conference on Neural Networks*, volume of Plenary, Panel and Special Sessions, pages 94-99, Washington, D.C.

Prodromidis, A., Chan, P., & Stolfo, S. (2000). Meta-Learning in Distributed Data Mining Systems: Issues and Approaches. In Kargupta H. & Chan P. K. (Eds.), *Advances in Distributed and Parallel Knowledge Discovery*. AAAI/MIT Press.

Pedrycz, W. (1997). Data Mining and Knowledge Discovery: A Fuzzy Set Perspective, *In Proceedings of the Seventh IFSA Congress*, pages 195-218, Czech Republic.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Shen Q. & Chouchoulas A. (1999). Combining Rough Sets and Data-Driven Fuzzy Learning for Generation of Classification Rules. *Pattern Recognition*, 32(12), pages 2073-2076.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8, pages 338-353.