

Future Directions for Machine Learning

J. F. Baldwin

Engineering Mathematics Department

University of Bristol

Email: jim.baldwin@bristol.ac.uk

Summary

In this paper we discuss possible future directions of research for soft computing in the context of artificial intelligence machine learning. Fundamental issues are presented with basic ideas emphasised rather than detailed accounts of algorithms and procedures.

The use of fuzzy sets for machine learning, computer intelligence and creativity are discussed in relation to the central problems of creating knowledge from data, pattern recognition, making summaries, user modelling for computer / human interfaces, co-operative learning, fuzzy inheritance for associative reasoning.

A voting model semantics for fuzzy sets is assumed with the corresponding ideas of mass assignment theory. The use of fuzzy sets in this way will provide better interpolation, greater knowledge compression, less dependence on the effects of noisy data than if only crisp sets were used. We will see how easy and useful it is to use successful inference methods such as decision trees, probabilistic fuzzy logic type rules, Bayesian nets with attributes taking fuzzy values rather than crisp values.

Introduction

As a society we collect so much data but do so little effectively with it. Supermarkets collect data concerning customer transactions, report distribution difficulties, and provide financial accounts. Hospitals carry out tests and hold patient records. Data is provided in engineering projects to provide models for analysis and design. Financial Institutions provide records of potential customers, collect statistics of existing customer performances. Traffic behaviour, market surveys, product performance are recorded to improve management and performance decisions. Data is cleaned and processed, studied and moved from place to place, visualised and communicated,

stored in data banks. We collect with enthusiasm hoping that the data will provide us with the answers we require. Data Mining is a new subject area with journals and Conferences, books describing case studies and recalling methods of analysing and visualising data. Methods are available for answering questions of a single source of data by forming rules to predict the answers, generalising from those cases in the database. Neural net methods, Bayesian nets and Decision trees are some of the more sophisticated approaches to provide these generalisations. So often background knowledge is ignored, applications are rather simple and constrained, the methods are not robust for noisy data and for data in which attributes are continuous variables. We can answer simple queries about one attribute when given values of some or all the others in the database. We can find simple patterns in the data. The tuples in the database are used as learning examples and how good the predictions and classifications are depend on how well this data is representative of the situation in hand.

Ideally we would wish to throw the data away and replace it with a summary in the form of knowledge sentences which fully represented all the relevant knowledge contained in the data. This summary would form part of a larger knowledge base made up of similar summaries from other database sources. We could then converse with the knowledge base in the same way as we might wish to talk to a human expert. The knowledge base could be used to provide deductive inferences, probabilistic inferences, analogical inferences, associative and inheritance inferences. We are far from this ideal but we need to move towards it.

The central question for this paper is to ask how can soft computing help and is it really necessary. I will ask the more specific question. How can fuzzy sets be of help in our pursuit for creating knowledge summaries from databases? This avoids the difficulty of deciphering what we actually mean by soft computing.

The world is not fuzzy in any way. We can look out and see the precision of a leaf on a tree, the curvatures of the road, the shading of a car, the detail of a person's face, the clouds in the sky. But this precision which we can see with our eyes if we want is often unwanted detail when it comes to labelling and categorising and classifying and clustering the real world into groups which we can label. We give labels to such objects as people, clouds, cars, fields, good paintings, happy faces etc. because we wish to talk about these objects in terms of their common properties within their group. It is a necessary part of our understanding of the world to label and give names to these labels and use them in natural language.

A crisp set is a set of elements which satisfy all our criteria for belonging to the set. An even set of dice values is $\{2, 4, 6\}$ because the numbers 2, 4, and 6 are all even and are dice values. A fuzzy set is a set of values with memberships in the range $[0, 1]$ where the membership values is in some sense a measure of how well the element satisfies what we mean by small dice value. A fuzzy set of small dice values might be $\{1 / 1, 2 / 0.8, 3 / 0.3\}$ where 2, for example has membership of 0.8 in the fuzzy set small associated with dice values. We will later discuss the semantics of this fuzzy set in more detail.

In moving from crisp to fuzzy sets we need only to introduce the membership function as having possible values in the range $[0, 1]$ instead of in the set $\{0, 1\}$ as for the crisp case. Most people would accept this change. The controversy of fuzzy sets has little to do with this definition. People might ask how can we choose the actual membership values and this is a sensible question which somehow we must answer. The controversy of fuzzy set theory lies with the calculus used for such operations as intersection, union etc. But there is really a more fundamental question which is asked before worrying about the calculus. Is it really necessary to complicate our modelling world by introducing fuzzy sets in addition to existing crisp sets? We will now discuss this important question.

Every one will accept that humans use fuzzy concepts. In fact, almost all our concepts are fuzzy. This contrasts with the mathematical world which uses precisely defined concepts. Why do we use fuzzy concepts, concepts that we cannot define

exactly. We have no definition for tall and yet we use the term and find it easy to apply. When is an object a bush rather than a tree? These questions have been asked many times and over several years. Do we really require discussing them now? We need to discuss them to answer why we require fuzzy sets for the purpose of converting large data sets into knowledge bases. We must look more closely at the semantics of a fuzzy set, study in more detail how we should modify those methods which use only crisp sets to accommodate fuzzy sets and indicate the real gain in using these more complicated ideas.

We will try to answer why we need fuzzy sets before discussing their semantics and what calculus we should use. This will provide motivation for studying them in more detail. Most instrument measurements provide us with exact numbers – these numbers may be approximate and would be better interpreted as a mean of some probability distribution. Attributes of databases mostly have exact values in their tuples. There may be missing information and sometimes only a range of possible values given. Sometimes the value may be described by a probability distribution. Only very infrequently would an attribute be given a fuzzy value. Why should we therefore bother with fuzzy sets?

Consider an attribute, which can take a continuum of values in some range. We would expect every object or data tuple which had attributes with similar values to be similar in respect to classification and predictions for any unknown attribute. We would like to group values of an attribute together and give them a label to distinguish this group from others. We could use precise groupings, intervals defining possible ranges of values. We could then group tuples together which had the attribute values in the same intervals. Tuples whose values belong to the same intervals would then be indistinguishable. If a certain classification were associated with one we would also associate it with the others. Similarly we would provide the same prediction for all. A tuple would belong to a cell in a multidimensional space. All tuples with the attribute values lying in the same intervals would belong to the same cell. A classification or prediction would be given to each cell. No account can now be given to the fact that one tuple may lie on the edge of one cell and near other cells. You would expect tuples near the edge of the cell to be a little similar to those

belonging to neighbouring cells. If we took this into account it would be equivalent to some form of multidimensional interpolation. This would obviously provide greater accuracy for prediction. In more complicated situations we would not have a definite classification associated with each cell but a probability distribution over the set of possible classifications. In this case the multidimensional interpolation would change the probability distributions of tuples belonging to the same cell by taking into account the distributions of the neighbouring cells. We see, therefore, that the multidimensional interpolation provides better accuracy for both prediction and classification.

The formation of cells and the grouping of tuples with the multidimensional form of interpolation provide a natural method of generalisation. We have classification and prediction for any point in a given cell even though the database only provides a selection of points in the cells.

How can we obtain the multidimensional interpolation easily and naturally? If we use fuzzy sets as defining our intervals rather than crisp ranges then we will have the required variation across the intervals by means of the membership function of the fuzzy set. Points in the centre of the interval will be given high membership and those at the edges low membership with intermediate values in between. Thus instead of dividing a line up into crisp intervals we can define the intervals using such fuzzy sets as **small**, **medium** and **large**. This will provide us with the required variation and interpolation effect. We will discuss exactly how in a later section.

This use of fuzzy sets is not restricted to attributes with continuous variables. The same arguments apply to the case where we have a discrete set of possible values for an attribute. For example we can replace the dice values {1, 2, 3, 4, 5, 6} with the groupings {**small**, **medium**, **large**} where **small**, **medium** and **large** are fuzzy sets defined in the space of dice values.

The fuzzy sets we will label as words. We can still use the methods we might use with crisply defined groupings with these fuzzy groupings. How we do this? What modifications must we make so that we can deal with fuzzy labels? We will discuss this in a later section. For now it is enough to say that this can easily be done.

The effect of using multidimensional interpolation not only increases the accuracy but will allow us to use fewer rules if we are using rules as our form of knowledge representation, smaller decision trees if this is the form of knowledge representation, or more simple graphs if we use Bayesian nets. The use of fuzzy sets to provide groupings for our data thus gives better accuracy, greater data compression and natural generalisation. This is our motivation for using fuzzy sets. The additional complication introduced by using fuzzy sets rather than crisp intervals pays off because of the greater accuracy and compression.

The labels on the groupings are words represented with fuzzy sets. A modified counting procedure of the mass assignment theory, Baldwin *** is used to convert the original database to a reduced database in which the attributes take the fuzzy set labels as values. We can then use classical machine learning methods such as ID3, Bayesian Nets and rule formation methods on this reduced database in the normal way. The approach to be discussed is also relevant to inductive logic programming methods with probabilistic conditioned rules in which predicate variables can be instantiated to fuzzy sets. The final classification will be a probabilistic distribution over the possible classifications.

The fusion problem is most important and has been discussed for several years. The following is an example of a fusion problem. We see colours in a three dimensional colour space. If we damage one of the dimensions we effectively project on to one of the three two dimensional spaces. There are three types of colour blindness corresponding to each of these two dimensional projections. Imagine we have three such colour blind persons viewing a colour scheme. Each provides a colour description from their two dimensional point of view. Can we construct a correct three dimensional view point from these descriptions?

This problem arises often in data mining because we have too many attributes, too many variables, too many dimensions to cope with. We can project onto lower dimensional spaces and do our classification with respect to each of these smaller dimensional points of view. We must then fuse the answers together to obtain the solution for the full dimensional space. The fusion can take various forms from simple conjunctive fusion, a weighted average sort of fusion to the more complex case in

which no simple formula based fusion can be assumed and we have to learn the rules of fusion for the case in hand from examples.

May be the most important problem we have is how to select the most appropriate attributes. The primary attributes of the database may not be the most appropriate and combinations of these primary attributes would form better discrimination. The combination can be algebraic, relational and combinations of these. How can we select these more appropriate attributes? We will suggest genetic programming where trees representing language constructs depicting relational and algebraic combinations are formed to satisfy some fitness measure. The choice of this fitness measure is most important. We cannot use the success of a construct on a test set because this would be computationally too severe. We need some form of discrimination measure that would indicate possible success. The mass assignment theory can supply such a measure.

We have indicated in this introduction that fuzzy sets can be used effectively to provide considerable enhancement to those methods that are already used but with crisp groupings. Because we are using fuzzy groupings it is important that necessary modifications to existing methods should be simple extensions. Changing from crisp to fuzzy intervals or groupings should not require radical rethinking or radically different methods. Methods we use with the fuzzy groupings will always have their counterpart with crisp groupings. All we wish for in using fuzzy groupings is that we obtain better accuracy and greater compression.

Two Viewpoints for Machine Learning

Consider a classification problem in which objects with attributes $\{A_1, \dots, A_n\}$ are to be classified with class labels $\{C_1, \dots, C_m\}$. A training set of examples Tr is given and also a Test set Ts . The attribute domains are covered with words given by trapezoidal or triangular fuzzy sets as described previous publications. Each object in Tr of a given class C_i can be allocated as a distribution over the word cells of the multidimensional attribute word space. The object will have values for each of the attributes. These values are distributed over the set of words associated with that space. Suppose that associated with the domain of attribute A_k we have the set of words $\{w_{ki}\}$ then the value of A_k will give a distribution over $\{w_{ki}\}$. This will be the case whether the value of the attribute is a precise

value or fuzzy set defined on the domain of A_k . We repeat this for values of each of the attributes. We multiply these distributions to obtain a distribution over the word space associated with $A_1 \times A_2 \dots \times A_n$, B say. This provides a count of the proportions of this object in the multidimensional word cells. We repeat with all the examples in Tr for class C_k . If the example has a fuzzy representation for its classification over the set of class labels then this fuzzy set is converted to a distribution over the classes and that proportion belonging to class k is distributed over the word space as just described. We can therefore deal with examples with fuzzy classifications as well as pure classifications. This is important for prediction type problems in which the classes are fuzzy sets on the prediction space. A predicted value for an example in Tr would have a fuzzy set representation over the class labels.

The examples in Tr provides us with a distribution over B for each class. We have effectively $Pr(C_k | \text{multidimensional word cell})$ for all k and all word cells. These distributions can be represented as extended Fril rules, Baldwin 19986, 1993, 1996, Baldwin, Martin, Pilsworth 1995, 1998.

Alternatively, the distribution over the multidimensional word space can be converted to a discrete fuzzy set F as described in previous publications and the rule for classification C_k is

Classification is C_k
IFF attribute word values belong to F

This is an equivalence rule which can be expressed directly in Fril.

These two viewpoints are equivalent from Fril point of view since the inference methods of Fril are based on probability theory and the mass assignment theory is used to give this interpretation. For example, if for a new example, it has a distribution over B which has a fuzzy set representation of F' , then $Pr(F | F')$ is determined using semantic unification. Both F and F' are discrete fuzzy sets defined over B . This represents the probability of the body being true. This probability is passed to the head of the rule since we are using an equivalence rule.

The same result is obtained from the extended Fril rule.

In the case of prediction, the classes C_k will be words on the output space. The extended Fril rules, one for each classification, will, for a new case, provide a probability distribution, $\{p_k\}$ over the $\{C_k\}$. The defuzzified predicted point value is then the weighted sum $\sum_k p_k v_k$ where v_k is the expected value of the least prejudiced distribution of the fuzzy set associated with the word representing class C_k .

This has no difficulties for low dimensional problems. But we have the problem of the curse of dimensionality. As we have more attributes to consider we have an exponential growth in the computation. We can get the most efficient extended Fril rule by using a mass assignment ID3 approach as discussed below although this more efficient decision tree may introduce a bias as far as the generalisation to examples in T_x is concerned. This is the case because some branches in the decision tree are not present in order to represent the correct classifications in the training set T_r . These might though be of value for cases in T_x . This is less likely to occur when using fuzzy sets than for the case when only crisp sets are used.

We deal with the curse of dimensionality by breaking the problem into sub-problems of lower dimension containing a sub-group of attributes.. The solution of the lower dimensional problems are then fused together. The most extreme form is to consider each of the attributes separately. Thus in this approach we find a classification rule with respect to each attribute, namely

Class is C_k from attribute A_i point of view IFF attribute A_i belongs to F_{ki} where F_{ki} is a discrete fuzzy set over the words associated with attribute A_i .

We could also use the conditional probabilities $Pr(w_{ki} | \text{class } k)$ in an extended Fril rule

Class is C_k from attribute A_i point of view
with $\{p_k\}$ probability $Pr(w_{ki} | \text{class } K)$
IFF attribute A_k is w_{ki} }_{all k}

Thus for a given new example we will use these rules to determine a distribution over the classes from each attribute point of view.

How do we fuse these view points together? Two obvious schemes are apparent.

(1) Conjunctive fusion

In this case we multiply the probabilities of the various view points for a given classification. This is equivalent to using the conjunctive fuzzy rule

Class is C_k IFF attribute A_1 belongs to F_{k1} and attribute A_2 belongs to F_{k2} and ... Attribute A_n belongs to F_{kn} .

(2) Evidential fusion

In this case we take a weighted average of the supports for class k from the different view points. This is equivalent to using the Fril evidential logic rule:

Class is C_k
IFF A_1 belongs to F_{k1} with weight w_{k1}
 A_2 belongs to F_{k2} with weight w_{k2}
.
.
.
An belongs to F_{kn} with weight w_{kn}
The weights are optimised.

These two approaches with each attribute taken separately was used as the basic method in the Fril data browser. It has some relationship to using naïve Bayes for classification.

We can extend these approaches to higher dimensional sub-sets than that of single dimensions. The problem is what sub-sets should we choose. We can use overlapping subsets of attributes. Groups of attributes can be chosen and each one used separately. The solutions are fused as above for the one dimensional case. Ad hoc methods which have some intuitive appeal have been used to select groupings of attributes for this purpose.

Mass Assignment ID3, Belief Networks and other applications

Both ID3 and Bayesian Net techniques can be used directly on the reduced database formed using fuzzy set labels on the attribute spaces and the modified counting procedure.

In this way the decision tree is grown and the only difference from the normal ID3 method, Quinlan 1990, 1995. is the use of the modified counting procedure to determine the proportions of tuples passed along each branch. This added computation is trivial and, in fact, because fewer branches will generally be required when using the fuzzy

divisions of the continuous variables or fuzzy groupings of discrete variables the overall computation can be reduced from the normal ID3 approach. Both pre and post pruning methods can be used to obtain an efficient decision tree. This can then be converted directly to a Fril extended rule for each class.

The class distribution for a new case can be determined by using these Fril extended rules, one for each class and the inference methods of Fril

The reduced database can also be used to find the conditional probabilities for Bayesian belief network architectures. The variables in a belief network can take a discrete set of values. We can deal with continuous variables by allowing the variable to take word values expressed by fuzzy sets. Also discrete variables can have word values corresponding to fuzzy sets defined on the set of values of the discrete sets. This extension to the fuzzy case will allow for more simple computations to determine the joint distribution since we are finding the joint distribution over the multidimensional word space. Variables will take values corresponding to these word labels. Once the conditional probability tables have been constructed using the modified counting procedure above the same approaches for inference can be used as is used at present with Belief Networks, [Pearl 1984, Jensen 1996]. The arguments for the use of fuzzy defined words rather than crisp intervals or exact groupings are as we have already explained for classification. We get better interpolation between similar cases and reduced computation. The Bayesian Net architecture assumes certain conditional independencies and this provides added efficiency over say ID3 when these independencies are ignored.

This Net approach can be used to determine the probability distribution for any unknown variable value when given values for the other variables.

Both these approaches have been used for function approximation and other data mining type problems and give very accurate results.

Conclusions

We have indicated many important areas of application of the modified counting procedure to machine learning. The modified counting procedure allows the transformation from the original attribute cross product space to a new

multidimensional word space in which each word is a fuzzy set. This transformation allows continuous and discrete variables to be represented by fuzzy labels which we have called words. The use of fuzzy words gives greater accuracy because natural interpolation is invoked, greater compression because of the more efficient representation, and a natural smoothing of noisy data. Other applications to intelligent systems along similar lines can be created. This provides an exciting future for the use of fuzzy sets in the field of artificial intelligence. There is no conflict with the probabilistic approach, no conflict with existing acceptable methods.

References

Baldwin, J. F., Martin, T. P. and Pilsworth, B. W. (1995). "FRIL - Fuzzy and Evidential Reasoning in AI", Research Studies Press (John Wiley).

Baldwin, J. F. (1996). "Knowledge from Data using Fril and Fuzzy Methods" in Fuzzy Logic in AI, Ed. J. F. Baldwin, John Wiley. 33-76.

Baldwin, J. F. and Martin, T. P. (1997). "Basic Concepts of a Fuzzy Logic Data Browser with Applications" in Software Agents and Soft Computing: Concepts and Applications, Ed. H. S. Nwana and N. Azarmi, Springer (LNAI 1198). 211-241.

Jensen F., (1996), An Introduction to Bayesian Networks Springer-Verlag.

Pearl, J., (1988), Probabilistic reasoning in Intelligent Systems : Networks of plausible inference, Morgan Kaufmann