Type-2 Fuzzy Classifier Ensembles for Text Entailment

Asli Celikyilmaz[1] I. Burhan Turksen[2]

[1]Electrical Eng.& Computer Sciences Dept., Univ. of California, Berkeley, CA, USA
[2]Dept. of Industrial Eng., TOBB Economy and Technology University Ankara, Turkey

Abstract

This paper presents a new Type-2 Fuzzy Classifier ensemble, which enables to model parameter uncertainties by characterizing the fuzzy sets with secondary membership values. We use fuzzy clustering method to characterize primary membership values and genetic algorithm to approximate secondary membership grades. Furthermore, a weighing algorithm is used for a non-complex reduction for reasoning. We use transductive reasoning, instead of inductive reasoning, to develop a local model for every new vector, based on a nearness criterion vectors from the given database. It is shown that the method can improve classifier system modeling performance in comparison to well-known methods.

Keywords: type-2 fuzzy sets, classifier ensembles, fuzzy c-classification.

1. Introduction

The concept of *type-2 fuzzy set* (T2FS) was introduced by Zadeh [1] as an extension of type-1 fuzzy set to identify the uncertainties present in fuzzy systems. With fuzzy sets of higher type (e.g. type-2), the fuzziness of relations is increased to handle inexact information. T2FSs are useful in situations, when it is difficult or uncertain to determine the exact MF of a fuzzy set, primary MFs [2], [3], [4]. In such cases, interval T2FS are defined (Fig. 1) which identify the footprint-of-uncertainty (FOU).

Despite its success in modeling real systems under uncertainty [2]-[6], implementation of type-2 fuzzy systems is not as easy as type-1 fuzzy systems due to complicated operations of T2FSs, mainly type-reduction. In addition characterizing secondary membership grades is a difficult task.

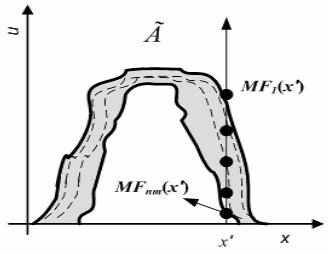

Fig. 1: Footprint-of-uncertainty of Fuzzy Set. Each membership function (MF) value is characterized with embedded type-1 fuzzy set.

To overcome some of the challenges of type-2 fuzzy system computations, in this paper, we propose a practical approach for classification problems. We initially build fuzzy classifier ensembles (multiple classifiers) by fuzzy partitioning the given dataset using a Fuzzy *c*-Classifier (FCC) method and obtain as many discriminant functions for each partition. The level of fuzziness parameter, *m*, of fuzzy clustering methods, which determine the degree of overlap of clusters, viz. structures, granules, etc., is used in many different research to identify the FOU of MFs [2],[5]-[6]. In an analogical manner, we identify the FOU of MFs using the FCC for discrete values of the fuzziness

parameter $m^r>1$, $r=1..nm$ and identify as many discriminant functions $f_{i,nm}(x)$, $i=1\ldots c$: for each cluster i. Each discrete m^r characterizes a fuzzy classifier model and identifies the interval valued MFs for each ensemble model.

We identify the optimum secondary MF grades, i.e., weights, of the primary MF grades obtained from the FCC models using genetic algorithms. New data vectors adopt the secondary MF grades obtained from the training samples in their neighborhood. During genetic learning process, each individual in the population encodes these weights for each training vector for each cluster, separately. This is quite a cumbersome process when the number of training vectors is large; therefore we implemented the transductive learning method [12]. Instead of learning the secondary MF grades of the entire training dataset, a new set of weights are learnt for each new data point from fairly few training vectors, which are in vicinity of the corresponding new vector.

We applied the new fuzzy modeling tool as the answer selection module of the whole Question and Answering (QA) system[7], in which the aim is to find precise answers to natural language questions from large document collections. We want to retrieve candidate answers and rank them based on a *textual entailment* model. An entailment relation between two text snippets (text-hypothesis pair) is produced when the hypothesis' meaning can be inferred from the text's.

We first convert the question query into a regular sentence (hypothesis-*h*) and then use *textual entailment* module to identify if the candidate sentence (text-*t*) entails *h.* Given the text and hypothesis:

t:Harry was born in Iowa.

h:Harry's birthplace is Iowa.

t entails *h,* otherwise we recognize the relation between the meaning of the texts as false entailment. We implement the proposed type-2 fuzzy classifier ensemble to build an entailment module for Question/Answering system and show that it could be an alternative method to well-known classifier methods.

2. Type-2 Fuzzy Classifier Ensembles

The T2FC is a type-2 fuzzy inference system akin to Takagi-Sugeno type inference systems and yet identifies one membership function for the entire antecedent part. We assume that the membership functions of each input variable are not independent, and their interactive affect should be analyzed instead of their individual effect. The secondary membership values are optimized with genetic algorithms. The first step of T2FC is to fuzzy partition the entire dataset into overlapping classifiers using the Fuzzy C-Classifier (FCC) algorithm akin to the fuzzy c-regression clustering method [8].

Let f_i be a function of nv dimensional feature vectors $\mathbf{x}_k(x_{k,1}\ldots x_{k,nv})\in X,\ k=1,..,n$ data points with binary class labels, $l(\mathbf{x}_k)\in\{0,1\}$. Each cluster $i=1\ldots c$ is represented with a discriminant function by:

$$p_{i,k}(l(\mathbf{x_k})\mid \mathbf{x_k}) = \mathbf{e}^{f_i(x_k)}/\mathbf{1}+\mathbf{e}^{f_i(x_k)},$$
$$f_i(\mathbf{x_k}) = \beta_{0,i} + \textstyle\sum_j \beta_{j,i} x_{j,k}, \quad i=1\ldots c \qquad (1)$$

The $p_{i,k}(l(x_k)=1|x_k)$ is the posterior probability for class $l(x_k)=1$ given x_k.

***Step* 1**: Assign c-classifier functions f_i as initial cluster representatives. For each iteration t:

Step* 2:* Calculate the $c\times n$ membership matrix $u_{i,k}\in U^{(t)}$, $u_{i,k}\in[0,1]$ as follows:

$$E_{i,k} = \left(l(\mathbf{x_k}) - p_{i,k}(l(\mathbf{x_k})\mid \mathbf{x_k})\right)^2,\ 1\le i\le c$$
$$u_{i,k}^{(t)} = \left(\sum_{j=1}^{c}\left(\frac{E_{i,k}}{E_{i,k}}\right)^{1/(m-1)}\right)^{-1} \ for\ \sum_{i=1}^{c} u_{i,k}^{(t)} = 1 \qquad (2)$$

The closer posterior probability to the actual class label, the less error will be.

Step* 3:* If $\|U^{(t)}-U^{(t-1)}\|\le\varepsilon$, then stop; otherwise go to step 4.

***Step* 4**: Using the *weighted logistic regression* method, calculate the new clus-

ter representatives for the (t+1)th iteration, $\beta_i^{(t+1)}=(x^T w_i x)' x^T w_i adjy$, where w_i denote the diagonal matrix of $\Re^{n\times n}$ having $u_{k,i}^{(t)} \in U_i^{(t)}$ as kth diagonal elements. The $adjy = (xw_i)+[(y-f_i)/f_i']$ is the Taylor expansion of the log-likelihood of the posterior probabilities $p_{i,k}$ [9].

3.1. T2FC Secondary MF grades

The membership values $u_{i,k}$ in (2) depends on the level of fuzziness parameter, $m \in (1,\infty)$, which determines the fuzziness of the resulting clusters. T2FC performs the following learning algorithm:

Execute FCC Metdod. To identify FOU of T2FS, the FCC is executed for different *levels of fuzziness,* $m^r=\{m^1 \ldots m^r\}$, $r=1 \ldots nm$, given the number of clusters, c. Each FCC model is characterized with each discrete m^r value and identifies classifier function $f_i^r(\mathbf{x},\beta_i^r)$ for each cluster, $i=1 \ldots c$ to obtain membership function values $MF_i^r(x)=u_i^r(x)$ and posterior probabilities $p_i^r(l(x_k)|x_k,\beta_i^r)$.

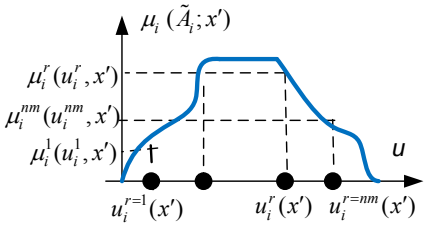

Fig. 2: Type-2 Fuzzy set --Secondary membership values of x' in cluster i based on each discrete fuzziness parameter m^r.

Initialize Secondary T2FSs. Each possible discrete membership value $u_i^r(x')$ $r=1 \ldots nm$ are randomly assigned initial weights, viz., secondary MF grades $\mu_{i,k}^r(\tilde{A};x_k) \in [0,1]$, $k=1 \ldots n$. (Fig. 2), since we don't have prior information what their values would be beforehand. These MF grades denote possibilities associated with each m^r at each value of x, $x_k=x'$.

Genetic Learning Process (GLP). Optimum values of the secondary MF grades of T2FSs at each x_k is identified based on genetic learning process. At this point, transductive learning algorithm is implemented to estimate the secondary membership values of data vectors.

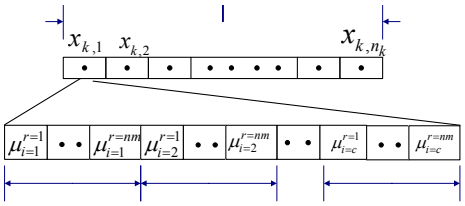

Fig. 3: A chromosome for $x_k=x'$.

When a new vector x' is introduced, a new model is build to estimate its output. The secondary MFs of x' in each cluster is estimated using n_k nearest neighbors from training dataset, which form a sample dataset $x_j \in X_j=\{x_1 \ldots x_{nk}\}$. Each chromosome of x' is encoded using initial weights of each n_k training vectors, one for each discrete m^r value for each cluster. In Fig. 3 $x_{k,j}$, $j=1..n_k$, represents each nearest vector to the selected $x_k=x'$ vector and $T2FS_i$ represents T2FS of the jth nearest train-vector in cluster $i=1..c$. Each $T2FS_i$ is identified by set of $u_{i,j}^r(x_j)$'s calculated by each m^r (Fig. 3). Herein, a genetic algorithm is used to optimize the secondary MF grades of nearest n_k vectors instead of the entire training dataset. A separate genetic learning method is executed for each new x' as follows:

Step-1 Initialize each chromosome in the population randomly and start iterating;

Step-2(i) Update secondary MF grades of each chromosome (Fig. 3) using mutation and crossover operations.

Step-2.(ii) For each chromosome, *chr*, calculate weighted posterior probabilities of each nearest train vector $x_{k,j}$ as follows:

$$\hat{p}_{j,chr} = \sum_{i=1}^{c} \frac{\sum_{r=1}^{nm} p_i^r(l(x_j)\,|\,x_j,\beta_i^r) u_{i,j}^r(x_j) \mu_{i,j}^r}{\sum_{r=1}^{nm} \mu_{i,j}^r} \qquad (4)$$

Step-2.(iii) Calculate performance index (PI) of each chromosome from each nearest training vector $x_{k,j}, j=1..n_k$

$$PI_{chr} = -\frac{1}{n_k}\sum_{j=1}^{n_k}(y_j - p_{j,chr})^2 \qquad (5)$$

***Step*-2.(*iv*)** Choose surviving individuals based on (arg_{chr} *max* PI_{chr}) and go to *Step*-2.(*i*) if termination condition is not satisfied which is either when the total number of iterations is reached or when there is no change in performances.

During GLP *nm* different secondary MF grades $\mu_{i,j}^{r}$ are identified for each discrete primary MF grade $u_{i,j}^{r}(x_j)$, $r=1...nm$, of each nearest training vector x_j of x'.

3.2. GT2FI Reasoning

We use the weighing formula of equation (4), to estimate the posterior probability of a particular vector x' using T2FC system. Firstly, the primary MF grades, $u_i^r(x')$ for each m^r value is calculated using equation (2). Since we do not know the actual label of x', we use actual class labels of the vectors in the vicinity of x', x_j. $j=1\ldots n_k$. To find the secondary MF grades, the weights of these nearest training vectors obtained from the GLP step are used. To calculate $u_i^r(x')$ for each $m=m^r$ using (2), we need the error values which we have no prior information, therefore, we use the error values of each x_j in each local model $\square_{i,j}^{r}(x_j)$. The secondary MF grades of nearest train vectors obtained from GLP are used to calculate one posterior probability value p_j' for the x' using (4). The implication and aggregation operators are combined in one step and thus the type of the MF is reduced down to type-1 first by using model weights captured in GLP step and then the fuzzy output probability $p_{i,j}$ is further reduced down to type-0 to obtain a single possibility value $p_j(x')$ using each nearest vector, x_j. To calculate a single crisp probability value for x', $p(x')$ the posterior probabilities of the nearest training points x_j $j=1..n_k$, are weighed based on inverse distance between x'. A sample output of the new T2FC using an artificial dataset is shown in Fig. 4.

3. Experiments on Text Entailment

We applied the proposed T2FC method on Textual Entailment datasets (freely available from PASCAL recognizing textual entailment (RTE) conference). The goal is to recognize semantic inference that a textual entailment defines directional relation between two text fragments, called *text* (*T*) and *hypothesis* (*H*) so that a human being can infer that *H* is most likely true on the basis of *T*.

3.1. Dataset

We combined different RTE datasets and only used the *T-H* pairs that are specifically designed for QA tasks. We extracted different sets of attributes from the *T-H* pairs (see Table 1) and to generate some of these features, we used different tools including Stanford Tagger, Named Entity Tagger, WordNet::Similarity Package.

Each (T-H) pair is analyzed to extract the features which depend on the relation between them. Some of these features are:

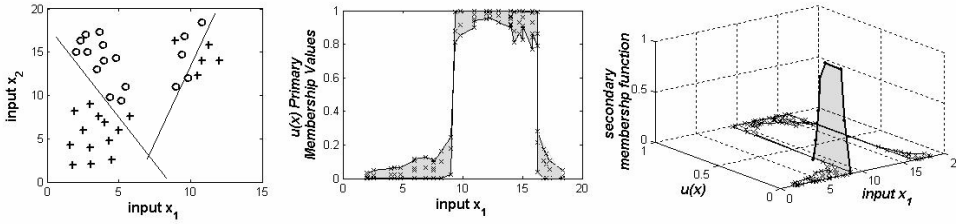

Fig. 4: (left) Artificial Dataset, (middle) FOU by $m\in[1.1, 2.6]$, (right) secondary MF of $x'=8.4$.

Table 1: Examples of Text-Hypothesis pairs from Recognizing Text Entailment Challenge.

Example Pairs
False Entailment **T**: In February 2002, President Bush visited China to mark the 30th anniversary of Nixon's historic trip. **H**: Nixon visited China in February 2002.
True Entailment **T**: Chernobyl nuclear-power plant is in Ukraine, but the reactor that exploded during the night of April 26, 1986, is 10 miles from the Belarusian border. **H**: The Chernobyl disaster took place on the 26th of April, 1986.

Lexico-Syntactic Overlap-Alignment Features*:* These features range from the ratio of the consecutive word overlap between the T and H (*n-gram, i.e.,* $n \in \{1,2,3\}$), the lowest common subsequence which measures the similarity between text T with length *m* and hypothesis H with length *n,* by searching in-sequence matches that reflect sentence level word order. We extracted these features for words and word-phrases, which are compounds of words.

Semantic Features: Noun, verb and adjective/adverb specific semantic overlap metric (similarity measure) using WordNet hypernym, hyponym, negation match between T-H based on clue phrases such as no-not neither, etc. are some of the examples of the features extracted from T-H pairs.

We created train and testing datasets using the T-H pairs from RTE challenge and extracted features as explained above which forms the inputs and for the binary output variable, 1 for "true entailment" and 0 for "false entailment" are assigned. We extracted 29 features using different combinations of the above features.

There were 2167 T-H pairs for building the learning models--training and 2400 pairs separated for testing purposes. Only 484 of the training pairs are QA based T-H pairs and consequently 526 of the testing ones are only created from QA based approaches. False and true entailments are evenly distributed.

3.2. Model Construction

The system model performance is measured with accuracy. To analyze the performance of the new system, the accuracy results of T2FC models are compared to well-known Adaptive Network Based Fuzzy Inference System (ANFIS) [10] which represents a hybrid type-1 fuzzy inference system, and support vector machines (SVM) [11] for classification, which is commonly used to build text entailment classification models.

Table 2: Parameters of benchmark tools.

Opt. Parameters
ANFIS: Hybrid method to optimize inference parameters, Gaussian MFs, TSK rule base structure.
SVM-LIN*:* $Creg \in [2^{-3},2^{7}]$, Linear kernel function, $K(x_k,x_j)= x_k{}^T x_k$
SVM-RBF*:* $Creg \in [2^{-3},2^{7}]$, Non-linear Gaussian radial basis kernel, $K(x_k,x_j)=\exp(-\gamma\|x_k-x_j\|)$, $\gamma>0$

For the FCC clustering of T2FCC, we set the boundaries of the level of fuzziness parameter between m_{lower}=1.4 and m_{upper}=2.6, which was proven to be min-max boundaries of the level of fuzziness parameter of the fuzzy *c*-models [15]. The *m* interval is discretisized into 10 values. The secondary MF values of nearest data points are optimized with genetic algorithms. For the genetic learning process, the initial population size and number of iterations are set to 100 each, and the number of clusters is set to 3. The crossover rate is set at 0.8 and the mutation rate is set at 0.01. Tournament selection with eliticist strategy is employed. The learning parameters of the rest of the methods are shown in Table 2.

The accuracy results of the experiments are shown in Table 3. The highest accuracy is obtained with the proposed T2FC method with 9% improvement on the testing cases. Since the T2FC is based on transductive learning, where a separate model for each testing case is build using

the nearing training cases for learning, training accuracy is not measured.

Table 3: Accuracy results of the Text Entailment for QA tasks.

Model	*Train Dataset*	*Testing Dataset*
ANFIS	0.694	0.500
SVM-LIN	0.655	0.549
SVM-RBF	0.647	0.555
T2FC	N/A	0.607

On of the challenges of T2FC is that the reasoning takes a longer time compared to the rest of the models since, for T2FC, for each new observed data, a new model is build. Hence, in the future we plan to build offline models as a consequence of the T2FC by using weighted models obtained from different training cases. Textual entailment task is a challenging problem and the accuracy of the outcome can be improved should other state-of-the-art NLP tools and semantic approach are used. This is left out as a future study.

4. Conclusions

In this paper, a type-2 fuzzy classifier ensemble system is introduced for binary classification domains. Unlike counterparts, local structures are characterized with discriminant functions to identify multiple-overlapping classifier models within the given structure. The uncertainty interval of primary membership functions (MF) are defined based on upper and lower limits of the level of fuzziness parameter of fuzzy c-classification method. The secondary MF grades are optimized with genetic algorithms. With the implementation of transductive learning method, a new model is constructed with only the training vectors in the vicinity of each new test vector. The algorithm adopts simple type-reduction and does not require defuzzification.

5. References

[1] L.A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences* vol. 8, pp. 199-249, 1975.

[2] I.B. Turksen, "Type-1 and Type II Fuzzy System Modeling," *Fuzzy Sets and Sys.* vol. 106, pp. 11-34, 1999.

[3] J. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems:Intr. And New Directions.* NJ: Prentice-Hall., 2001

[4] R. John and S Coupland, "Geometric type-1 and type-2 fuzzy logic systems," *IEEE Trans. Fuzzy Syst.* vol. 15, pp. 3-15, 2007.

[5] C. Hwang, F.C.-H. Rhee, "Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means," *IEEE Trans. Fuzzy Syst.* vol. 15, pp 107-120, 2007.

[6] A. Celikyilmaz and I.B. Turksen, "Uncertainty Modeling of Improved Fuzzy Functions with Evolutionary Systems," *IEEE Trans. Systems, Man and Cybern.* Vol. 38(4), 2008.

[7] P. Prager, E. Brown, A. Coden, D. Radev, "Question answering by predictive annotation," *Proc SIGIR,* 2000.

[8] R. Hathaway and J. Bezdek, "Switching regression model and fuzzy clustering," *IEEE Trans. Fuzzy Syst.* vol. 1, pp. 195-204, 1993.

[9] M. Jorgensen, "Iteratively reweighted least squares," Encyc. Of Environments, 2006.

[10] J-S.R. Jang (1993) ANFIS: Adaptive Network Based Fuzzy Inference System. *IEEE Trans. Syst., Man, Cybrn.* volume 23, pages 665-685.

[11] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[12] T. Joachims, "Transductive inference for text classification using support vector machines," *Int. Conf. Machine Learning* (ICML) pp. 200-209, 1999.

[13] A. Celikyilmaz, I. B. Turksen, "Uncertainty Bounds of Fuzzy C-Regression Method," *2008 IEEE Int.*

Conf. on Fuzzy Systems (WCCI-2008), pp. 1193-1198, 2008.