The Occam's Razor Principle for Data Mining Models Based on Degenerate Selfguessing Fuzzy Classification Algorithms

Sergey Vatlin

The National Bank of the Republic of Belarus. Settlement Centre. 7, Kalvariyskaya street, 200048, Minsk Republic of Belarus, vatlin_si@nbrb.by

Abstract: *Fuzzy classification models are one of the basic types of data mining models. The concepts of the simplicity and efficiency for fuzzy classifiers are introduced. We also introduced the concepts of consistent and degenerate selfguesssing fuzzy classifiers. The Occam's Razor principle for data mining models based on fuzzy classification algorithms is formulated. The quality criterion for degenerate selfguessing fuzzy classifiers based on invariant simplicity measure is proved. The theorems on the conditions of improvement of degenerate selfguessing fuzzy classifiers are proved.*

***Keywords*:** data mining, degenerate selfguessing fuzzy classifiers, the Occam's Razor principle for data mining models, simplicity measure, improvement of fuzzy classifiers.

1. INTRODUCTION

Classification models for transformation of information (classifiers) are one of the basic types of data mining models [1]. On the basis of such models a researcher can determine with the objects of which set (from a prefixed collection of sets L = (w_1 , w_2 , … ,w_L)) he operates at a fixed moment of time.

A comprehensive analysis of a broad range of existing classification models is presented in [2].

One of the possible directions in the development of classification models is connected with the theory of fuzzy sets. As L. Zadeh notes, the essential connection between such models and the fuzzy set theory is based on the fact that most of the real classes are fuzzy by their nature, i.e., transition from membership to not membership to these classes is rather gradual than discontinuous [3].

2. THE OCCAM'S RAZOR PRINCIPLE FOR DATA MINING MODELS BASED ON FUZZY CLASSIFICATION ALGORITHMS

Let **X (Card X = m)** and **L** (Card **L** << **m**) be the fixed sets of objects of unspecified nature. Let's denote be means of **Y** a family of all possible continuous functions translating **L** into a real-valued interval **[0.1] (μ ∈ Y ⇔ μ**: **L → [0.1]** and **μ** is a continuous function). The sets **X** and **Y** will be called further as sets of initial and finite symbols in the problem of fuzzy classification [4].

Let **f** be a target function from **X** to **Y, f: X→ Y**. Let's also denote by **Θ** a collection of ordered pairs from **X** × **Y** such that $\mathbf{\Theta} = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbf{X}$, $y_i = \mathbf{f}(x_i)$, $y_i \in \mathbf{Y}$, $\forall\, i = \overline{1.n}$. The set **Θ** will be referred further as a learning set (for a function **f)** and a number **n -** as a power of a learning set **Θ**.

Let **G** be an arbitrary fuzzy classification algorithm to take an **Θ** into a hypothesis function $\mathbf{h}_{G\theta}$ from **X** to **Y,**

$$\mathbf{h}_{G\theta} = \mathbf{G}(\mathbf{\Theta}),\ \mathbf{h}_{G\theta} : \mathbf{X} \to \mathbf{Y}.$$

The construction process of an acceptable hypothesis function (model) is usually performed within some linguistic structure, which provides a symbolic representation of the set of potential classification models **H**. Let's denote a set of all fuzzy classifiers, induced a family models (functions) **H** by **{G(Θ)}.** Let's also define a fixed language **Z** as a pair **(I,T)** where **T** is a set of sentences in the language, and **I** is interpreter **I**:**T→H**.

On each t∈ **T** we define a complexity measure **C** : **T→R** (for the fixed language **Z = (I,T)**) which may represent either syntactic or semantic aspects of a sentence t.

Under the fixed measure **C** we define the complexity **C**(**h**) for an arbitrary classification model(function) $\mathbf{h}_{G\theta}$ by the following rule:

$$\mathbf{C}(\mathbf{h}) = \frac{\min C(t)}{t : I(t) = h} \qquad (1)$$

In other words the complexity of classification model **h** depends on the language **Z =** (**I,T**) and is defined as the complexity of the simplest sentence which induced that model.

Let **Z =** (**I,T**) be a fixed language, **C** : **T→R –** a fixed complexity measure (for a language **Z)**, **X** and **Y –** sets of initial and finite symbols in the fixed fuzzy classification (data mining) problem.

Let also **f** correspond to a fixed (but unknown) target classification function and $\mathbf{\Theta} = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbf{X}$, $y_i = \mathbf{f}(x_i)$, $y_i \in \mathbf{Y}$, $\forall\, i = \overline{1.n}$ - a learning set for this function of n power.

Then, the Occam's Razor principle for data mining models based on fuzzy classification model may be formulated as follows:

Under otherwise equal conditions, favour such the element $\mathbf{G}^* \in$ **{G(Θ)}** that induces the simplest (in accordance the complexity measure (1)) fuzzy classification model $\mathbf{h}^* = \mathbf{h}_{G^*\Theta}$, that

$$\mathbf{G}^* \Leftrightarrow \frac{\min\{C(h_{G\Theta})\}}{G \in \{G(\Theta)\}}$$

The main problem that arises when using this formulation is the ambiguity problem of simplicity (complexity) identification of a fuzzy classification model **h**∈ **H**. The model that is simple in one linguistic structure may be a complicated one in another such structure.

But, at least for one family of fuzzy classifiers the Occam's Razor principle for data mining models based on such classification algorithms may be formulated by rigorously unambiguous way independently of the used language.

3. THE QUALITY CRITERION FOR DEGENERATE SELFGUESSING FUZZY CLASSIFIERS BASED ON INVARIANT SIMPLICITY MEASURE

Let $\Theta = \{(x_i, y_i)\}_{i=1}^{n}$ - be a learning set on **n** power for a target classification function **f: X→ Y, G –** a fuzzy classifier to take **Θ** into a hypothesis classification function $\mathbf{h}_{G\theta}$ from **X** to **Y,** $\mathbf{h}_{G\theta}$ **= G(Θ),** $\mathbf{h}_{G\theta}$ **: X→ Y.**

We'll say that **G** is consistent with a learning set **Θ iff** the following relationship is obeyed:

$$\mathbf{h}_{G\Theta}(\mathbf{x}_i) = \mathbf{y}_i, \forall\, i = \overline{1.n}.$$

Under fixed **X, Y** and **Θ** let's denote a set of all possible fuzzy classifiers, consistent with **Θ** by **Con(Θ)**.

Let **G** be a fixed element from **Con(Θ).**We'll say that **G** is a degenerate selfguessing for **Θ** [5] **iff:**

A) **G** is consistent with any subset $\mathbf{\Theta}^* \subset$ **Θ,** such that **Card** $\mathbf{\Theta}^* = \mathbf{n}^* > \mathbf{n}_0 - 1$ ($\mathbf{n}_0$ - a fixed parameter, characterizing the **X** set structure. In the case $\mathbf{X} = \mathbf{R}^m$, $\mathbf{n}_0 = \mathbf{m}+1$).

B) There exist a subset $\mathbf{\Theta}' \subset$ **Θ, Card** $\mathbf{\Theta}' = \mathbf{n}' \geq \mathbf{n}_0 - 1$ that is a function of **G** and **Θ,** $\mathbf{\Theta}' = \mathbf{\Phi}(\mathbf{G}, \mathbf{\Theta})$ so that:

B.1) $\mathbf{h}_{G\Theta'}(\mathbf{x}) = \mathbf{h}_{G\Theta}(\mathbf{x}),\ \forall\, \mathbf{x} \in \mathbf{\Theta} \setminus \mathbf{\Theta}'$;

B.2) if $(\mathbf{x}', \mathbf{y}') \notin \mathbf{\Theta}'$, then

$$\mathbf{\Phi}(\mathbf{G}, \mathbf{\Theta}) = \mathbf{\Phi}(\mathbf{G}, \mathbf{\Theta} \setminus (\mathbf{x}', \mathbf{y}')).$$

Under fixed **X, Y** and **Θ** let's denote a set of all possible fuzzy classifiers, degenerate selfguessing for **Θ** by **DSg(Θ)**.

Let $\mathbf{\Theta}_1$ (**Card** $\mathbf{\Theta}_1 = \mathbf{n}_1$) and $\mathbf{\Theta}_2 = \mathbf{\Theta}_1 \cup \{(\bar{x}, \bar{y})\}$ (**Card** $\mathbf{\Theta}_2 = \mathbf{n}_2 = \mathbf{n}_1 + 1$) be an arbitrary learning sets for a target classification function **f, f: X→ Y.**

Let's assume that in $\mathbf{Con}(\mathbf{\Theta}_1)$ there are, at least, two different elements $\mathbf{G}_1$ and $\mathbf{G}_2$.We'll say that a fuzzy classifier $\mathbf{G}_2$ is more effective than a fuzzy classifier $\mathbf{G}_1$ (while moving from learning set $\mathbf{\Theta}_1$ to learning set $\mathbf{\Theta}_2$) and put down this fact as $\mathbf{G}_2 \succ \mathbf{G}_1$ **iff** the following relationship is obeyed:

$$\mathbf{P}(\mathbf{h}_{G_2\Theta_1}(\bar{x}) = \bar{y}) > \mathbf{P}(\mathbf{h}_{G_1\Theta_1}(\bar{x}) = \bar{y}), \qquad (2)$$

where P(•) is the probability of event (•).

Let's also suppose that in $\mathbf{DSg}(\mathbf{\Theta}_1)$ there are, at least, two different elements $\mathbf{G}_1$ and $\mathbf{G}_2$ and denote:

$$\mathbf{n}_1' = \mathbf{Card}\,\mathbf{\Theta}_1',\ \mathbf{\Theta}_1' = \mathbf{\Phi}(\mathbf{G}, \mathbf{\Theta}_1);$$

$$\mathbf{n}_2' = \mathbf{Card}\,\Theta_2',\ \Theta_2' = \mathbf{\Phi}(\mathbf{G}, \mathbf{\Theta}_2).$$

Then, takes place

Theorem1 [5]

The probability of event $\mathbf{h}_{G_2\Theta_1}(\bar{x}) = \bar{y}$ is greater or equal than $\mathbf{1} - \frac{n_2'}{n_1 - 1}$

$$\mathbf{P}(\mathbf{h}_{G_2\Theta_1}(\bar{x}) = \bar{y}) \geq \mathbf{1} - \frac{n_2'}{n_1 - 1}.$$

We'll say that a degenerate selfguessing fuzzy classifier $\mathbf{G}_2$ is simpler than degenerate selfguessing fuzzy classifier $\mathbf{G}_1$ (while moving from learning set $\mathbf{\Theta}_1$ to learning set $\mathbf{\Theta}_2$) and put down this fact as $\mathbf{G}_2 \rhd \mathbf{G}_1$ **iff** the following relationship is obeyed:

$$\mathbf{n}_2' < \mathbf{n}_1' \qquad (3)$$

Takes place

Theorem2 (The quality criterion for degenerate selfguessing fuzzy classifiers based on invariant simplicity measure)

$$\mathbf{G}_2 \rhd \mathbf{G}_1 \Leftrightarrow \mathbf{G}_2 \succ \mathbf{G}_1.$$

Proof is directly follows from definitions (2), (3) and Theorem 1.

4. IMPROVEMENT OF THE CONSISTENTLY AND DEGENERATE SELFGUESSING FUZZY CLASSIFIERS

For an arbitrary learning set and fixed **X** and **Y** let's denote by $\{G_{XY}(\Theta)\}$ a set of all possible fuzzy classifiers, operating with **Θ** (We'll omit lower indexes and write down **{G(Θ)}** in the situations when it's obvious what **X** and **Y** we speak about).

Takes place

Theorem3

$$\mathbf{DSg}(\mathbf{\Theta}) \subset \mathbf{Con}(\Theta) \subset \{\mathbf{G}(\mathbf{\Theta})\}. \qquad (4)$$

Proof is directly follows from corresponding definitions.

Let's denote by F_Θ an arbitrary mapping of {G (Θ)} into itself (F_Θ :{G(Θ)} →{G(Θ)}) and by {L(G(Θ))} – a collection of all possible mappings F_Θ.

Let also K = { $\bar{\Theta}$ } corresponds to a fixed collection of learning sets from $2^{X \times Y}$ and Θ_1, $\Theta_2 = \Theta_1 \cup \{(\bar{x}, \bar{y})\}$ – to the arbitrary elements from K.

Let's denote by means of $\mathbf{G}_1$ and $\mathbf{G}_2$ the fixed element from $\{\mathbf{G}(\Theta_1)\}$**.** We'll say that $\mathbf{G}_2$ **improves** $\mathbf{G}_1$ on the element $\bar{x} \in$ **X**, and put down this fact as

$$\mathbf{G}_2 = \bar{F}_{\Theta_1}(\mathbf{G}_1) \text{ iff}$$

$$\mathbf{P}(\mathbf{h}_{G_2\Theta_1}(\bar{x}) = \mathbf{f}(\bar{x})) > \mathbf{P}(\mathbf{h}_{G_1\Theta_1}(\bar{x}) = \mathbf{f}(\bar{x})).$$

Under fixed **Θ** and **G** let's denote by means of **{H(G(Θ))}** a set of all possible mappings $\bar{F}_\Theta$.

Let **G** be a fixed fuzzy classifier and $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2 = \mathbf{\Theta}_1 \cup \{(\bar{x}, \bar{y})\}$ such learning sets that $\mathbf{G} \in \mathbf{DSg}(\mathbf{\Theta}_1) \cap \mathbf{DSg}(\mathbf{\Theta}_2)$.Then, takes place

Theorem4

For an arbitrary fuzzy classifier $\mathbf{G}_1 \in \mathbf{Con}(\mathbf{\Theta}_1) \setminus \mathbf{DSg}(\mathbf{\Theta}_2)$ exist $\bar{F}_{\Theta_1} \in \{\mathbf{H}(\mathbf{G}_1(\mathbf{\Theta}_1))\}$ such that $\mathbf{G} = \bar{F}_{\Theta_1}(\mathbf{G}_1)$.

5. CONCLUSION

It was previously established [4-5] that in solving many practical and model machine learning, data mining and pattern recognition problems the efficiency of the degenerate selfguessing fuzzy classifiers is much higher than the efficiency of fuzzy classifiers which don't belong to **DSg(Θ).**

However, due to the famous induction paradoxes [6], such behavior of the degenerate selfguessing fuzzy classifiers can't be universal.

That is why it is important to elucidate conditions under which the joint use of degenerate selfguessing and improvement principles will bring us to guaranteed results.

One of such conditions is given Theorem 4 revealing restrictions under which a degenerate fuzzy classifier could be an improvement of a consistently fuzzy classifiers.

Theorem 3 shows that pay for the mentioned improvements may be high since the move from the right to the left in inclusion (4) cannot be performed automatically.

Fuzzy classifiers are heuristic algorithms that transform information. The heuristic nature of these algorithms is determined by the principle boundedness of the power of learning set Θ.

The models of reasoning, embedded in each heuristic, specify a confidence level of the corresponding fuzzy classifier, this level may be deviated far from 100%. In such a situation one possible way for increasing the efficiency of fuzzy classifiers is connected with the postulation of an informal principal for selection the "best" fuzzy classifier among the set of alternative, roughly equivalent ones.

One of the principles is the Occam's Razor principle that acquired a good reputation itself in the physical science. According to this principle when other conditions being equal we must chose the simplest model among two different ones.

Unfortunately, when employing this principle in data mining practice the identification problem of the model's simplicity (complexity) arises. The data mining (or machine learning or pattern recognition) model that is simple in one linguistic structure may be a complicated one in another such structure.

The first effort to investigate the possibility of an operational interpretation of the Occam's Razor principle in terms of data mining (machine learning) models based on fuzzy classifiers seems to be undertaken in the paper.

It is shown that at least for one family of such classifiers this principle may be formulated by a rigorously unambiguous way irrespective of the used simplicity measure and linguistic structure.

Theorem 2 gives the quality criterion for degenerate selfguessing fuzzy classifiers based on invariant simplicity measure.

6. REFERENCES

[1] D.E. Holmes (Editor), L.C. Jain (Editor). Data mining: Foundations and intelligent paradigms: Volume 1: Clustering, association and classification. Springer. Berlin, 2011.p.352.

[2] R.S. Michalski, J.G. Carbonnel, T.M. Mitchell. Machine learning: An artificial learning approach. M. Kaufmann. Palo Alto, CA, 1983. p.482.

[3] L. Zadeh. Fuzzy sets and their application to pattern classification and cluster analysis. In J. van Ryzin (Editor). Classification and Clustering. Academic Press. New York, 1977. pp. 148-223.

[4] S.I.Vatlin. The design of fuzzy classifiers on the basis of two level correction method, *Electromagnetic Waves and Electronic Systems* 6(1997), p. 32-37.

[5] S.I.Vatlin. Strongly selfguessing fuzzy classifiers, *Informatica (Vilnius)* 6 (1995), p. 85-93.

[6] N. Goodman Fact, fiction and forecast. Harvard University Press. Cambridge, 1975. p. 246.