Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 7/2 (2014), 101-107
DOI: http://dx.doi.org/10.21609/jiki.v7i2.263

SIMILARITY BASED ENTROPY ON FEATURE SELECTION FOR HIGH DIMENSIONAL DATA CLASSIFICATION

Jayanti Yusmah Sari, Mutmainnah Muchtar, Mohammad Zarkasi, dan Agus Zainal Arifin

Department of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Jl. Teknik Kimia, Kampus ITS Sukolilo, Surabaya, 60111

E-mail: jayanti13@mhs.if.its.ac.id, agusza@cs.its.ac.id

Abstract

Curse of dimensionality is a major problem in most classification tasks. Feature transformation and feature selection as a feature reduction method can be applied to overcome this problem. Despite of its good performance, feature transformation is not easily interpretable because the physical meaning of the original features cannot be retrieved. On the other side, feature selection with its simple computational process is able to reduce unwanted features and visualize the data to facilitate data understanding. We propose a new feature selection method using similarity based entropy to overcome the high dimensional data problem. Using 6 datasets with high dimensional feature, we computed the similarity between feature vector and class vector. Then we find the maximum similarity that can be used for calculating the entropy values of each feature. The selected features are features that having higher entropy than mean entropy of overall features. The fuzzy k-NN classifier was implemented to evaluate the selected features. The experiment result shows that proposed method is able to deal with high dimensional data problem with mean accuracy of 80.5%.

Keywords: *classification, entropy, feature selection, high dimensional data, similarity*

Abstrak

Curse of dimensionality merupakan masalah yang sering dihadapi pada proses klasifikasi. Transformasi fitur dan seleksi fitur sebagai metode dalam reduksi fitur bisa diterapkan untuk mengatasi masalah ini. Terlepas dari performanya yang baik, transformasi fitur sulit untuk diinterpretasikan karena ciri fisik dari fitur-fitur yang asli tidak dapat diperoleh kembali. Di sisi lain, seleksi fitur dengan proses komputasinya yang sederhana bisa mereduksi fitur-fitur yang tidak diperlukan dan mampu merepresentasikan data untuk memudahkan pemahaman terhadap data. Pada penelitian ini diajukan metode seleksi fitur baru yang berdasarkan pada dua pendekatan filter, yaitu *similarity* (kemiripan) dan entropi untuk mengatasi masalah data berdimensi tinggi. Tahap awal metode ini adalah menghitung nilai *similarity* antara fitur dengan vektor kelas dari 6 data berdimensi tinggi. Kemudian diperoleh nilai *similarity* maksimum yang digunakan untuk menghitung nilai entropi untuk setiap fitur. Fitur yang dipilih adalah fitur yang memiliki nilai entropi lebih tinggi daripada entropi rata-rata seluruh fitur. Fuzzy *k*-NN diterapkan untuk tahap klasifikasi data hasil seleksi fitur. Hasil percobaan menunjukkan bahwa metode yang diajukan mampu mengklasifikasi data berdimensi tinggi dengan rata-rata akurasi 80.5%.

Kata Kunci: *klasifikasi, entropi, seleksi fitur, data berdimensi tinggi, similarity*

1. Introduction

Curse of dimensionality with regard to the presence of large number of features is widely known as a major obstacle in classification task, because it is practically impossible to adequately populate the feature space with the available data. Reduction of feature dimensionality is considerably important to overcome this high dimensional data problem. The purpose of dimensionality reduction is to improve the classification performance through the removal of redundant or irrelevant features. Dimensionality reduction can be achieved in two different ways, which are feature transformation and feature selection. Feature transformation methods construct new features out of original variables and feature selection methods keep only useful features and discard others.

Feature transformation aims to build a new feature space of reduced dimensionality, produce a compact representation of the information that may be distributed across several of the original features. Ravi et al. [1] have developed an approach by using feature transformation method, called PCA-Ravi for deriving fuzzy rules to handle high-dimensional classification problems. Althou-

gh it has shown promising results in many applications [2], feature transformation is not easily to interpret because the physical meaning of the original features cannot be retrieved. On the other side, feature selection, as a preprocessing step to machine learning, is efective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [3].

Feature selection has important role in classification because it can simplify the model and make the model more transparent and more comprehensiv. There are three types of feature selection approaches: embedded, filters, and wrappers approaches. In embedded techniques, feature selection can be considered to be a part of the learning itself. By testing the values of certain features, algorithms split the training data into subsets. Filter techniques are designed to filter out undesirable features by checking data consistency and eliminating features whose information content is represented by others. The filter approach was also usually performs some statistical analysis without employing any learning model. Zhang et al. [4] have developed Constraint Score method. This is a filter method for feature selection with pairwise constraints, which specifies whether a pair of data samples belong to the same class (*must- link* constraints) or different classes (*cannot-link* constraint). Also, Luukka [5] has proposed a filter technique based on fuzzy entropy measures and tested it together with similarity classifier to do the feature selection in high dimensional medical datasets.

On the other hand, wrapper technique involves a learning model and uses its performance as the evaluation criterion. A research in wrapper technique was conducted by Aydogan et al. [6] which proposed a hybrid heuristic approach (called hGA) based on genetic algorithm (GA) and integer-programming formulation (IPF) to solve high dimensional classification problems in linguistic fuzzy rule-based classification systems. Tsakonas [7] has designed a genetic programming (GP)-based Fuzzy Rule Based Classification System as a learning process, called GP-PITT-Tsakonas, to generate complete fuzzy rule sets. Berlanga et al. [8] have proposed a GP-COACH method, a Genetic Programming-based method for the learning of COmpact and ACcurate fuzzy rule-based classifycation systems for high-dimensional problems. Although the wrapper approach is known to be more accurate compared to the filter approach [9,10]. But, it also tends to be more computationally expensive since the classifier must be trained for each candidate subset and do not scale up well to large, high-dimensional datasets.

In [5] all the features that having the higher entropy than the mean entropy are removed, but Jaganathan and Kuppuchamy [11] have stated that features with highest entropy values were the most informative ones. In this paper, we propose a new selection feature method based on two types of filter techniques, which are similarity and entropy measure. Our idea is to use the highest similarity as membership value in entropy measure and keep the features that have higher entropy than the mean entropy. This idea is proposed to overcome the complexity of learning algorithm while still preserving the good accuracy of the overall system, especially on high dimensional data classification. For classification step, we considered the fuzzy *k*-NN [12] to be a suitable classifier since it does not need any learning algorithm and the membership assignments to classify samples tend to possess desirable qualities.

2. Methods

Dataset

The datasets were downloaded from UCI Machine Learning Repository [13]. The fundamental properties of the datasets are shown in Table 1.

TABLE 1
DATASETS AND THEIR PROPERTIES

Datasets	Nb. of classes	Nb. of features	Nb. of instances
Ecoli	6	7	336
Glass	6	9	214
Heart D.	2	12	270
Ionosphere	2	34	351
Parkinsons	2	22	195
Wine	3	13	178

Ecoli
Ecoli dataset was created by Kenta Nakai from Osaka University. The dataset patterns were characterized by attributes calculated from the amino acid sequences. Each pattern has 7 attributes and 336 labeled examples.

Glass
The study of classification of types of glass was motivated by criminological investigation. Glass dataset was coming from USA Forensic Science Service, which contains 9 types of glass, defined in terms of their oxide content (Mg: Magnesium, Al: Aluminum, 6. Si: Silicon, etc.).

Heart Disease
The heart disease dataset was coming from V.A. Medical Center, Long Beach and Cleveland Clinic

Foundation. It composed of 297 measurements and 9 attributes. There are no missing values. The heart disease dataset includes 12 attributes and 2 classes.

Ionosphere
Ionosphere dataset was mostly used for classifycation of radar returns from the ionosphere. With 34 continuous attribute, this radar data was collected by a system in Goose Bay, Labrador. *"Good"* radar returns are those showing evidence of some type of structure in the ionosphere. *"Bad"* returns are those that do not; their signals pass through the ionosphere.

Parkinsons
The dataset was created by Max Little from the University of Oxford. Dataset is composed of a range of biomedical voice measurements from healthy people and people with Parkinsons disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from people who participated in collection of this data.

Wine
The dataset were the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 features found in each of the three types of wines. All attributes are continuous.

Data Preprocessing

All datasets have t number of different kinds of features (feature vectors) $f_1, \ldots, f_t$ and a label class (class vectors) $v_i = (v_i(f_1), \ldots, v_i(f_t))$. We suppose that the values for the magnitude of each attribute are normalized so that they can be presented as a value between 0 to 1. In order to attain this task, we should convert class vectors into a non-zero labeled class and normalize the feature vectors. Once the class vectors v has been converted, we used similarity and entropy measure to select the most informative features.

Similarity Measure

Let sample data $x = (x(f_1), \ldots, x(f_t))$, $\boldsymbol{x} \in X$, $f \in v$ in feature vector v. The decision to which class an arbitrarily chosen x belongs is made by comparing it to each class vector v. The comparison can be done by using similarity as given by equation(1) in the generalized Luka-siewicz structure [14]:

$$S(\boldsymbol{x},\boldsymbol{v}) = \left(\sum_{r=1}^{t} w_r\left(1 - |x(f_r)^p - v(f_r)^p|\right)\right)^{1/p} \quad (1)$$

for $\boldsymbol{x}, \boldsymbol{v} \in [0,1]$. Here, p is a parameter coming from the generalized Łukasiewicz structure [15] (p in (0, infinity) as default p=1) and w_r is a weight parameter, which set to one. If the sample belongs to class i, we get the similarity value between the class vector and sample being $S(\boldsymbol{x},\boldsymbol{v}) = 1$. If the sample does not belong to this class in class vector, we got 0 from the similarity value. The decision to which class the sample belongs was made according to which class vector the sample has the highest similarity value [5]. The similarity value is calculated using equation(2).

$$S(\boldsymbol{x}, \boldsymbol{v_i}) = \max_{i=1\ldots N} S(\boldsymbol{x}, \boldsymbol{v_i}) \quad (2)$$

This highest similarity was used as membership of $\boldsymbol{x}$, $\mu_A(x_j)$, for calculating its entropy.

Entropy Measure

We calculated the fuzzy entropy values for each features by using similarity values between the class vectors and feature vectors we want to classify. Entropy is a measure of the amount of uncertainty in the outcome of a random experiment, or equivalently, a measure of the information obtained when the outcome is observed.

De Luca [16] suggested the formula to measure fuzzy entropy that corresponded to concept of fuzzy sets and Shannon probabilistic entropy [17] in the following equation(3)

$$H(A) = -\sum_{j=1}^{n}(\mu_A(x_j)\log \mu_A(x_j) + (1-\mu_A(x_j))\log(1-\mu_A(x_j)) \quad (3)$$

where $H(A)$ is the measure of fuzzy entropy and $\mu_A(x_j)$ is the maximum similarity from the previous step, similarity measure. This fuzzy entropy measure was used to calculate the relevance of the features in feature selection process.

Feature Selection

We used the maximum similarity value from similarity measure as entropy value $\mu_A(x_j)$ of each feature. The highest fuzzy entropy value of the feature is regarded as the most informative one [14]. A feature $f \in F$ is selected if it satisfies the following condition of Mean Selection (MS) Strategy as shown by equation(4).

$$\sigma(f) \geq \sum_{f \in F} \frac{\sigma(f)}{|F|} \quad (4)$$

where $\sigma(f)$ is the relevance value of the features, that is selected if it is greater than or equivalent to the mean of the relevant values. This strategy will be useful in examining the suitability of the fuzzy

```
Step 1: Initialize classvec[1,...,l],
     Datalearn[1,...,m],
     feature[1,...,t]
Step 2: Compute the similarity between
     feature vector and class vector
     sim[j][i][k]=(1-classvec[j][i][k]^p
                  -Datalearn[i][j]^p)^(1/p)
Step 3: Sort similarity values
     sim[i][j][k] and find the maximum
     similarity value, max(S(x,v))=μi(xj)
Step 4: Compute entropy of each feature
        H[i] = - ∑x∈U μi(x)logμi(x) +
               (1 − μi(x)) log μi(x)
Step 5: Compute mean entropy,
     entropy_avg = sum(H)/t
Step 6: Remove feature which have
     entropy lower than entropy_avg
```

Figure 1. The algorithm of proposed feature selection method

entropy relevance measure. Our proposed feature selection method is presented in Figure 1.

In the algorithm, we have *m* samples, *t* features and *l* classes. We measure the similarities between feature vectors and class vectors using equation(1), and find the entropy value of each feature using equation(3). The mean entropy from all of entropy values is then calculated. The feature whose entropy value is lower than mean entropy value is removed from the dataset while the feature with higher entropy value is selected and used for classification.

Fuzzy *k*-NN Classification

After selecting the best features *t*, we classify the sample dataset *x* using fuzzy *k*-NN classifier. The basic concept of this classifier is to assign membership as a function of the object's distance from its *k*-nearest neighbors and the memberships in the possible class *l*. The pseudo-code of fuzzy *k*-NN classifier is presented in Figure 2.

Consider $W=\{w_1, w_2, ..., w_m\}$ a set of *m* labeled data, *x* is the input for classification, *k* is the number of closest neighbors of *x* and *E* is the set of *k* nearest neighbors (NN). Let $\mu_i(x)$ is the membership of *x* in the class *i*, *m* be the number of elements that identify the classes *l*, and *W* be the set that contain the *m* elements. To calculate $\mu_i(x)$, we use equation(5) [12].

$$\mu_i(x) = \frac{\sum_{j=1}^{k} \mu_{ij}\left(\frac{1}{\|x-m_j\|^{2/(m-1)}}\right)}{\sum_{j=1}^{k}\left(\frac{1}{\|x-m_j\|^{2/(m-1)}}\right)} \qquad (5)$$

Since we use fuzzy *k*-NN method, each element of *x* testing data is classified in more than one class with membership value $\mu_i(x)$. The decision to which class the elemen of *x* testing data belongs is made according to which class the element of *x* testing data has the highest membership value $\mu_i(x)$. Figure 3 shows the overall steps of our proposed method.

```
Set k
{Calculating the NN}
  for i = 1 to t
     Calculate distance from x to mi
     if i<=k
       then add mi to E
     else if mi is closer to x than any
          previous NN
       then delete the farthest neighbor
          and include mi in the set E
```

Figure 2. Pseudo-code of fuzzy *k*-NN classifier.

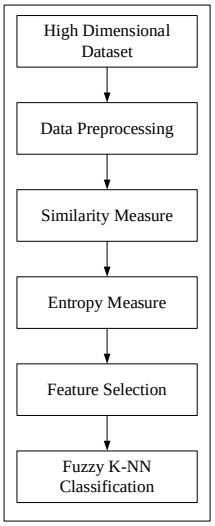

Figure 3. Proposed method's scheme.

3. Results and Analysis

The experiment was conducted to prove that feature selection method using similarity and entropy, can be used to classify high dimensional data. The performances of the proposed method were evaluated using 10-fold cross validation. All datasets were split into 10 data subsets. One subset was used for testing and the other nine subsets were used as sample. This procedure was repeated 10 times for all of datasets. The sample subset was used to select feature based on similarity and entropy value, while the testing subset is applied to evaluate the feature selected that obtained by the proposed method. The result of feature selection in Table 2 shows all the selected features from six different datasets. The lowest proportion of selected feature belongs to Ionosphere dataset and the highest one belongs to E-coli dataset.

The classification results are presented as accuracy. Accuracy is a comparison between the nu-

TABLE 2
FEATURE SELECTION RESULT

Dataset	Nb. of Feature	Nb. of Selected Feature	Selected Feature
Ecoli	7	5	1, 2, 5, 6, 7
Glass	9	5	3, 4, 6, 8, 9
Heart Disease	12	6	2, 3, 6, 7, 9, 12
Ionosphere	34	16	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32
Parkinsons	22	11	1, 3, 9, 10, 11, 12, 14, 17, 19, 20, 22
Wine	13	7	1, 2, 6, 7, 8, 10, 12

TABLE 3
CLASSIFICATION RESULT

Dataset	Nb.of *k*	Accuracy (%)	
		Lower Entropy	Higher Entropy
Ecoli	6	43.2	**84.0**
Glass	6	60.4	**62.4**
Heart Disease	2	57.8	**73.7**
Ionosphere	2	83.8	**83.3**
Parkinsons	2	72.6	**83.0**
Wine	3	80.3	**96.6**

mber of correctly classified data and misclassified data. Accuracy is calculated using equation(6).

$$Accuracy = \frac{n.correctly\ classified\ data}{n.misclassified\ data}\ x\ 100\% \quad (6)$$

Since we use 10-fold cross-validation procedure, the predictive accuracies on the testing set of the 10 runs of each dataset is averaged and reported as the predictive accuracies. In Table 3, classification results with predictive accuracy are reported for all of the datasets. To prove that feature with higher entropy value is the most informative features, we also performed an experiment using lower entropy value as a comparison.

Table 3 shows the performance of feature selection methods for classification using fuzzy *k*-NN classifier. The second column shows the number of *k* used for each dataset which is equal to the total of its classes. The third column is the result of feature selection method by using entropy values lower than the mean entropy. The fourth column is the result of feature selection method by using entropy values higher than the mean entropy value. The highest classification result obtained from Wine dataset with 96.6% of accuracy, while the lowest is obtained by Glass dataset with 62.4 % of accuracy.

The proposed feature selection method has reduced the number of features instead of using

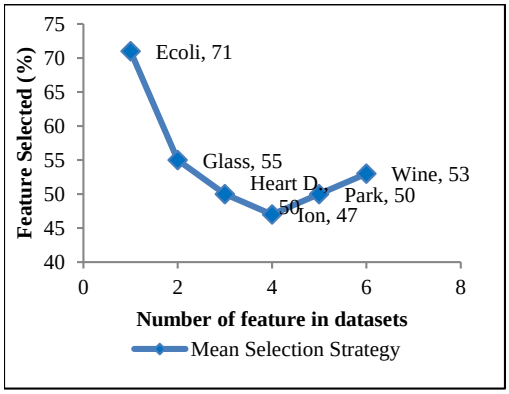

Figure 4. Feature selection result.

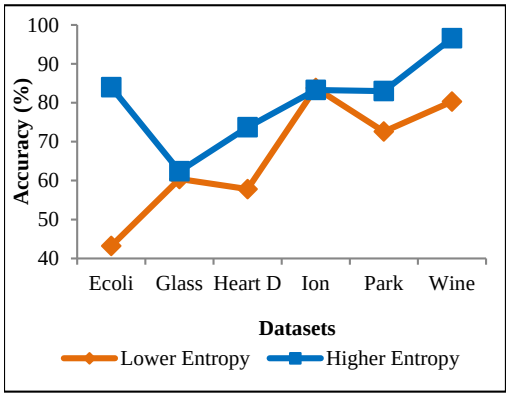

Figure 5. Comparison of accuracy.

all the features to perform the classification. The number of features (x-axis) in Mean Selection (MS) strategy is plotted against the percentage of features selected (y-axis) in the dataset in Figure 4.

Most of the datasets are getting the selected features approximately half of their features (54.3 %). This is because our proposed method implements the Mean Selection strategy which selecting the features with entropy values greater than or equivalent to the mean of the relevant values. The remaining features that cannot satisfy this threshold were then ignored. Except for E-coli dataset, we get more than half of the overall features (71%). This is because the E-coli dataset is having the largest class compared to other datasets. Large class tends to create the smaller similarity between features. Small similarity then leads to high entropy value that causes too many features to be selected.

As can be seen in Table 3, the features selected that coming from below the mean entropy value are having lower accuracy than the features from above the mean value. We get the 80.5% of mean classification accuracy by choosing features that having entropy above the mean entropy, whi-

TABEL 4
COMPARISON OF CLASSIFICATION RESULT

Dataset	[1]	[7]	[8]	[4]	Proposed Method
Ecoli	55.46	43.94	77.72	-	**84.02**
Glass	46.56	45.12	**65.33**	-	62.40
Heart D	49.24	56.46	55.23	-	**73.70**
Ion	-	-	-	82.20	**83.29**
Park	73.63	74.53	**86.48**	-	82.92
Wine	93.17	38.19	95.10	71.10	**96.59**

le choosing lower one lead to 66.35% of mean accuracy.

Figure 5 shows the comparison between these two conditions in term of accuracy. It proves that higher fuzzy entropy value of the feature is regarded as the most informative one.

We have compared our proposed method results with other previous works in feature reduction of high dimensional data [1,4,7,8]. Table 4 shows the comparison of classification accuracy of our proposed method to other methods. As can be seen, our proposed method has the highest accuracy in four datasets, which are E-coli (84.02 %), Heart Disease (73.70%), Ionosphere (83.29%) and Wine (96.59%) while a better result in Glass and Parkinsons datasets is showed from method [16]. Although the [16] was better than our proposed method when applied to the two datasets, it does not have a high difference in classification result. Glass in [16] shows 65.53%, while our method is having 62.4% of accuracy. Also, on Parkinsons dataset, [16] shows 86.48% of accuracy, whereas 82.92% of accuracy is derived from our method.

In our proposed method, best result is found in Wine dataset with 96.6% of classification accuracy and the lowest accuracy is shown in Glass dataset with 62.4%. This is because the Glass dataset is having a high number of classes, 7 classes, while our proposed method is better applied on data with high number of features, not classes. From this result, our proposed method manages to perform better than other previous researches using the same datasets, with its advantages in term of its simple feature selection and classification method.

4. Conclusion

We have presented a method for feature selection using similarity based entropy. The experiment was conducted by using 6 high dimensional datasets taken from UCI Machine Learning Repository. Similarity measure was implemented to find the similarity value between particular features to its class. This similarity value was used as membership for calculating the entropy of each feature. Features having entropy higher than the mean entropy are then selected. Result shows that the proposed method is more accurate compared to some methods proposed in previous works. Our proposed method is also able to handle the high dimensionality problem in term of the number of features, but not classes. In the future, this method can be considered as a promising feature selection method especially if combined with other feature selection methods, to overcome the high dimensionality problem in term of high features and high classes.

References

[1] V. Ravi, P. Reddy, H. Zimmermann, "Pattern classification with principal component analysis and fuzzy rule bases", *European Journal of Operational Research*, Vol. 126, pp 526–533. 2000.

[2] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization", *Journal of Machine Learning Research*, vol. 3, pp 1415-1438. 2003.

[3] A.Tsanas, M.A. Little, P.E. McSharry, "A Simple Filter Benchmark for Feature Selection", 2010, unpublished.

[4] D. Zhang, S. Chen, Z. Zhou, "Constraint Score: A New Filter Method for Feature Selection with Pairwise Constraints", *Pattern Recognition*, vol. 41, pp 1440-1451. 2007.

[5] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier", *Expert Systems with Applications*, vol. 38, pp 4600–4607. 2011.

[6] E.K. Aydogan, I. Karaoglan, P.M. Pardalos, "hGA: Hybrid Genetic Algorithm in Fuzzy Rule-Based Classification Systems for High-Dimensional Problems", *Applied Soft Computing*, vol. 12, pp 800–806. 2011.

[7] A. Tsakonas, "A Comparison Of Classification Accuracy Of Four Genetic Programming-Evolved Intelligent Structures", *Information Sciences*, vol. 176, pp 691–724. 2005.

[8] F.J. Berlanga, A.J. Rivera, M.J. del Jesus, & F. Herrera, "GP-COACH: Genetic Programming-based learning of COmpact and ACcurate fuzzy rule-based classification systems for High-dimensional problems", *Information Sciences*, vol. 180, pp 1183 1200. 2009.

[9] R. Kohavi, G.H. John, "Wrappers for Feature Subset Selection", *Artificial Intelligence*, vol. 97, pp 273–324. 1997.

[10] M.M. Kabir, M.M. Islam, K. Murase, "A New Wrapper Feature Selection Approach Using Neural Network", *Neuro computing*, vol. 73, pp 3273–3283. 2010.

[11] P. Jaganathan, R. Kuppuchamy, “A threshold fuzzy entropy based feature selection for medical”, *Computers in Biology and Medicine*, vol. 43, pp 2222-2229. 2013.

[12] J.E. Keller, M.R. Gray, J.A. Givens, “A fuzzy *k*-Nearest Neighbor Algorithm”, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, pp 580–585. 1985.

[13] UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/

[14] Łukasiewicz, J, Selected works, Cambridge University Press, 1970.

[15] Luukka, P., Saastamoinen, K., Kononen, V, “A classifier based on the maximal fuzzy similarity in the generalized Łukasiewicz-structure”, In Proceedings of the *FUZZY-IEEE 2001 conference*, Melbourne, Australia, pp. 195–198. 2001.

[16] De Luca, A., Termini, S, “A definition of non-probabilistic entropy in setting of fuzzy set theory”, *Information Control*, vol. 20, pp 301–312. 1971.

[17] Shannon, C. E, “A mathematical theory of communication”, *Bell System Technical Journal*, pp 379–423, 623–659. 1948.