Fuzzy Systems in Data Science and Big Data

Francisco Herrera
Soft Computing and Information Intelligent Systems (SCI²S)
University of Granada, Spain
Email: {herrera, alberto}@decsai.ugr.es
http://sci2s.ugr.es
Outline

- Introduction to Big Data. Big data analytics
- Fuzzy-based models for Big Data Learning
- Smart Data: The missing bridge between Big Data and real applications to get high quality data
- Fuzzy Big Data Science: Opportunities
- Final Comments

Thanks to Isaac Triguero and Alberto Fernandez for preparing part of this material
Outline

- Introduction to Big Data. Big data analytics
- Fuzzy-based models for Big Data Learning
- Smart Data: The missing bridge between Big Data and real applications to get high quality data
- Fuzzy Big Data Science: Opportunities
- Final Comments
Our world revolves around the data

- **Science**
  - Data bases from astronomy, genomics, environmental data, transportation data, ...

- **Humanities and Social Sciences**
  - Scanned books, historical documents, social interactions data, ...

- **Business & Commerce**
  - Corporate sales, stock market transactions, census, airline traffic, ...

- **Entertainment**
  - Internet images, Hollywood movies, MP3 files, ...

- **Medicine**
  - MRI & CT scans, patient records, ...

- **Industry, Energy, ...**
  - Sensors, ...
What is Big Data?

There is not a standard definition!

“Big Data” involves data whose volume, diversity and complexity requires new algorithms, technologies and analyses to extract valuable knowledge (hidden).
What is Big Data? The 5V’s definition

- Volume
- Velocity
- Variety
- Veracity

Value
What is Big Data? (in short)

Big data refers to any problem characteristic that represents a challenge to process it with traditional applications.
New programming model: **MapReduce**

- "Moving computation is cheaper than moving computation and data at the same time"

**Idea**

- Data is distributed among nodes (distributed file system)
- Functions/operations to process data are distributed to all the computing nodes
- Each computing node works with the data stored in it
- Only the necessary data is moved across the network
MapReduce: How it works

MapReduce algorithm [Dea04]

Google File System (GFS) [Ghe03]
Hadoop

- **Open source framework for Big Data processing**
  - **Based on** two Works published by Google
    - Google File System (GFS) [Ghe03]
    - MapReduce algorithm [Dea04]
  - **Composed of**
    - Hadoop Distributed File System (HDFS) → Storage
    - Implementation of the MapReduce algorithm → Processing

Hadoop

- Hadoop is:
  - An open-source framework written in Java
  - Distributed storage of very large data sets (Big Data)
  - Distributed processing of very large data sets

- This framework consists of a number of modules
  - Hadoop Common
  - Hadoop Distributed File System (HDFS)
  - Hadoop YARN – resource manager
  - Hadoop MapReduce – programming model

http://hadoop.apache.org/
Automatic parallelization:
- Depending on the size of the input data → there will be multiple MAP tasks!
- Depending on the number of Keys <k, value> → there will be multiple REDUCE tasks!

Scalability:
- It may work over every data center or cluster of computers.

Transparent for the programmer
- Fault-tolerant mechanism.
- Automatic communications among computers
Paradigms that do not fit with Hadoop MapReduce

- **Directed Acyclic Graph (DAG) model:**
  - The DAG defines the dataflow of the application, and the vertices of the graph defines the operations on the data

- **Graph model:**
  - More complex graph models that better represent the dataflow of the application
  - Cyclic models -> Iterativity.

- **Iterative MapReduce model:**
  - An extended programming model that supports iterative MapReduce computations efficiently
Big Data: Technology and Chronology

2001-2010

2010-2017

2001
3V’s Gartner
Doug Laney

2004
MapReduce
Google
Jeffrey Dean

2008
Hadoop
Yahoo!
Doug Cutting

2010-2017:
Big Data Analytics:
Mahout, MLLib, ...

Hadoop Ecosystem
Applications
New Technology
Take-home message so far

- We need new strategies to deal with big datasets
  - Choosing the right technology is like choosing the right data structure in a program.

- The world of big data is rapidly changing. Being up-to-date is difficult but necessary.
Big Data Analytics

Potential scenarios:

- Clustering
- Classification
- Recommendation Systems
- Social Media Mining
- Social Big Data
- Real Time Analytics/Big Data Streams

Clustering
Classification
Recommendation Systems
Social Media Mining
Social Big Data
Real Time Analytics/Big Data Streams
Data mining techniques have demonstrated to be very useful tools to extract new valuable knowledge from data.

The knowledge extraction process from big data has become a very difficult task for most of the classical and advanced data mining tools.

The main challenges are to deal with:

- The increasing scale of data
  - at the level of instances
  - at the level of features
- The complexity of the problem.
- And many other points (unstructured data, redundancy, ...)

Machine learning for Big Data
Spark and Flink Libraries

[Image: Diagram showing integration of Spark, Flink, and other libraries]

https://spark.apache.org/docs/latest/mllib-guide.html

Spark and Flink Libraries

[Diagram showing Spark Streaming, GraphX, Spark SQL, R, and Flink]

Spark and Flink Libraries

Spark Streaming

GraphX

Spark SQL

Spark Packages

A community index of third-party packages for Apache Spark.
354 packages

Spark and Flink Libraries

MLlib

FlinkML

Meta Store
HiveQL
UDFs
SerDes

Apache Spark

Spark Packages

A community index of third-party packages for Apache Spark.
354 packages
Outline

- Introduction to Big Data. Big data analytics
- Fuzzy-based models for Big Data Learning
- Smart Data: The missing bridge between Big Data and real applications to get high quality data
- Fuzzy Big Data Science: Opportunities
- Final Comments
Fuzzy-based models for Big Data learning

OUTLINE

- Fuzzy Rule Based Systems
- Data fragmentation and lack of data problem
- Big Data classification with fuzzy models
- Are FRBCSs robust with respect to the lack of data?
- Conclusions and future challenges
OUTLINE

- Fuzzy Rule Based Systems
- Data fragmentation and lack of data problem
- Big Data classification with fuzzy models
- Are FRBCSs robust with respect to the lack of data?
- Conclusions and future challenges
Fuzzy Rule Based Systems

- **Computational Intelligence** serves as the bridge to link natural intelligence and artificial intelligence for the scientific/engineering Application

  CI: Biologically and linguistically motivated computational paradigms

- The linguistic representation and processing power of fuzzy logic is a unique tool to bridge symbolic intelligence and numerical intelligence gracefully.

Source: CT Lin (Fuzz-IEEE, 2017)
Fuzzy Rule Based Systems

- **Machine learning:** The most effective classification algorithms do not allow to interpret the models.
- The interpretation of the models is an essential feature of rule-based systems (RBSs)
- There are applied areas where experts want to know why decisions:
  - Finances, medicine, security, ...
- RBSs fill a very important gap in Data Science: **INTERPRETATION.**
Fuzzy Rule Based Systems

- The most effective classification algorithms do not allow to interpret the models.
- The interpretation of the models is an essential feature of rule-based systems (RBSs).
- There are applied areas where experts want to know why decisions are made: Finance, Medicine, Security, ...
- RBSs fill a very important gap in Data Science: INTERPRETATION.

Fuzzy Rule Based Systems play an important role as RBSs for the description of the facts in terms of knowledge represented by rules:

INTERPRETATION
Fuzzy Rule Based Systems

**Deep Learning:** Transforms information to provide knowledge in terms of decisions, but do not explain why decisions (resurface neural networks).

**Fuzzy Rule Based Systems/Fuzzy Models:** Interpret information to provide knowledge in the form of linguistic rules or understandable and simple fuzzy models.

FRBSs should have a greater presence in the Data Science forums *(performing well, understandable and simple (N. Pal))*.
Fuzzy Rule Based Systems

Deep Learning Accuracy

Artificial Intelligence

Fuzzy Systems Interpretation
Fuzzy Rule Based Systems

Chi-FRBCS
Chi-FRBCS

- Generates rules as \textbf{“Rule R}_j\textbf{: IF } x_1 \text{ IS } A_1^j \text{ AND … AND } x_n \text{ IS } A_n^j \text{ THEN Class } = C_j \text{ with RW}_j\text{”}

- Builds the fuzzy partition using equally distributed triangular membership functions

- Builds the RB creating a fuzzy rule associated to each example

- Rules with the same antecedent may be created:
  - Same consequent \(\rightarrow\) Delete duplicated rules
  - Different consequent \(\rightarrow\) Preserve highest weight rule

\[\text{Z. Chi, H. Yan and T. Pham, Fuzzy algorithms with applications to image processing and pattern recognition, World Scientific, 1996.}\]
Rule Base (No Weights):
If \( x_1 \) is \textit{small} and \( x_2 \) is \textit{small} then Class 2
Fuzzy Rule Based Systems

Rule Base (No Weights)

If $x_1$ is small and $x_2$ is small then Class 2
If $x_1$ is small and $x_2$ is medium then Class 2
Fuzzy Rule Based Systems

Rule Base (No weights)

If $x_1$ is *small* and $x_2$ is *small* then Class 2

If $x_1$ is *small* and $x_2$ is *medium* then Class 2

If $x_1$ is *small* and $x_2$ is *large* then Class 1

1.0

large

medium

small

0.0

small

medium

large

$0.0 \quad x_1 \quad 1.0$
Fuzzy Rule Based Systems

Classification Boundaries

If $x_1$ is small and $x_2$ is small then Class 2
If $x_1$ is small and $x_2$ is medium then Class 2
If $x_1$ is small and $x_2$ is large then Class 1

... 

If $x_1$ is large and $x_2$ is large then Class 3

High interpretability
Standard rules (w/o weights) does not always get a good representation.

Rule Base (No Weights)

If \( x_1 \) is \textit{small} and \( x_2 \) is \textit{small} then Class 2

If \( x_1 \) is \textit{small} and \( x_2 \) is \textit{medium} then Class 2

If \( x_1 \) is \textit{small} and \( x_2 \) is \textit{large} then Class 1

\ldots

If \( x_1 \) is \textit{large} and \( x_2 \) is \textit{large} then Class 3

High Interpretability  
Low Accuracy
Fuzzy Rule Based Systems

RB (No weights)
If $x_1$ is small and $x_2$ is medium
then Class 2

RB (Weights == Certainty Factor)
If $x_1$ is small and $x_2$ is medium
then Class 2 with 0.75
There are other models beyond rule-based models (performing well, understandable and simple). **An interesting model:** fuzzy prototypes represented by nearest neighbor techniques.
Fuzzy-based models for Big Data learning

OUTLINE

- Fuzzy Rule Based Systems
- Data fragmentation and lack of data problem
- Big Data classification with fuzzy models
- Are FRBCSs robust with respect to the lack of data?
- Conclusions and future challenges
Rare cases or Small disjuncts are those disjuncts in the learned classifier that cover few training examples.
Big Data Classification: Data Fragmentation and Lack of Data Problem

It becomes very hard for the learning algorithm to obtain a model that is able to perform a good generalization when there is not enough data that represents the boundaries of the problem.

Minimize learning error + maximize generalization.
Big Data Classification: Data Fragmentation and Lack of Data Problem

**MapReduce**

Data Fragmentation with Parallel Processing + Model fusion

Small disjuncts arise with MapReduce data fragmentation. This problem is accentuated for imbalance classification: Lack of Data/lack of density between classes
Data fragmentation - Lack of data

The lack of data in the training data may cause the introduction of small disjuncts.

What it is also most significant, when the concentration of minority examples is so low that they can be simply treated as noise.
Big Data Classification: Data Fragmentation and Lack of Data Problem

Lack of data

Left-C4.5, right-Backpropagation (Pima data set): These results show that the performance of classifiers is repaired as the training set size increases. This suggests that small disjuncts play a role in the performance loss of class imbalanced domains.

Big Data Classification: Data Fragmentation and Lack of Data Problem

Lack of data. Fuzzy models performance

Robustness to the lack of data?
OUTLINE

- Fuzzy Rule Based Systems
- Data fragmentation and lack of data problem
- Big Data classification with fuzzy models
- Are FRBCSs robust with respect to the lack of data?
- Conclusions and future challenges
Uncertainty and Big Data

- Uncertainty is inherent to Big Data due to
  - Heterogeneous sources
  - Variety in data
  - Incomplete data
  - Veracity in question

- Fuzzy Rule Based Classification Systems can manage
  - Uncertainty
  - Vagueness
  - Lack of data/Data fragmentation
Linguistic FRBCSs have been used in distributed environments for a long time.

In the context of Big Data applications, MapReduce is the “key”:
- simple,
- fault-tolerant,
- scalable.

The programming framework differs from traditional schemes:
- Map tasks imply the building of local / partial models
- Reduce tasks aggregate partial models
Chi-FRBCS-BigData: A Case of Study

We choose a simple Learning Method to analyze the potential of FRBCSs for Big Data Classification

- MapReduce design based on the Chi et al. FRBCS algorithm (two different MapReduce processes)
  - Phase 1: Building the Fuzzy Rule Base
  - Phase 2: Estimating the class of samples belonging to Big Data sample sets
- Two versions that differ in the Reduce function of the building of the Fuzzy RB are considered
  - Chi-FRBCS-BigData-Max
  - Chi-FRBCS-BigData-Average

Chi-FRBCS

- Generates rules as “Rule R_j: IF x_1 IS A_{1,j} AND ... AND x_n IS A_{n,j} THEN Class = C_j with RW_j”
- Builds the fuzzy partition using equally distributed triangular membership functions
- Builds the RB creating a fuzzy rule associated to each example
- Rules with the same antecedent may be created:
  - Same consequent → Delete duplicated rules
  - Different consequent → Preserve highest weight rule

Z. Chi, H. Yan and T. Pham, Fuzzy algorithms with applications to image processing and pattern recognition, World Scientific, 1996.
Big Data Classification with Fuzzy Models

Building the RB with Chi-FRBCS-BigData: A Map Reduce approach

The key of a MapReduce data partitioning approach is usually on the reduce phase

Two alternative reducers (Max vs average weights)
Building the FRB with Chi-FRBCS-BigData-Max

\[ R_1: \text{IF } A_1 = L_1 \text{ AND } A_2 = L_1 \text{ THEN } C_1; \ RW_1 = 0.8743 \]
\[ R_2: \text{IF } A_1 = L_2 \text{ AND } A_2 = L_2 \text{ THEN } C_2; \ RW_2 = 0.9142 \]
\[ R_3: \text{IF } A_1 = L_1 \text{ AND } A_2 = L_2 \text{ THEN } C_2; \ RW_2 = 0.8842 \]
\[ R_4: \text{IF } A_1 = L_2 \text{ AND } A_2 = L_1 \text{ THEN } C_2; \ RW_3 = 0.6534 \]
\[ R_5: \text{IF } A_1 = L_3 \text{ AND } A_2 = L_2 \text{ THEN } C_2; \ RW_3 = 0.4715 \]
\[ R_6: \text{IF } A_1 = L_2 \text{ AND } A_2 = L_3 \text{ THEN } C_2; \ RW_3 = 0.7784 \]

\[ \text{REDUCE} \]

\[ R_1: \text{IF } A_1 = L_1 \text{ AND } A_2 = L_1 \text{ THEN } C_1; \ RW_1 = 0.9254 \]
\[ R_2: \text{IF } A_1 = L_2 \text{ AND } A_2 = L_2 \text{ THEN } C_2; \ RW_2 = 0.9142 \]
\[ R_3: \text{IF } A_1 = L_1 \text{ AND } A_2 = L_2 \text{ THEN } C_2; \ RW_2 = 0.8842 \]
\[ R_4: \text{IF } A_1 = L_2 \text{ AND } A_2 = L_1 \text{ THEN } C_2; \ RW_3 = 0.6534 \]
\[ R_5: \text{IF } A_1 = L_3 \text{ AND } A_2 = L_2 \text{ THEN } C_2; \ RW_3 = 0.4715 \]
\[ R_6: \text{IF } A_1 = L_2 \text{ AND } A_2 = L_3 \text{ THEN } C_2; \ RW_3 = 0.7784 \]

Final RB generation

\[ \text{RB}_1, R_1, C_1, RW = 0.8743 \]
\[ \text{RB}_2, R_1, C_2, RW = 0.9254 \]
\[ \text{RB}_3, R_2, C_1, RW = 0.7142 \]
\[ \text{RB}_4, R_1, C_2, RW = 0.2143 \]
\[ \text{RB}_5, R_2, C_1, RW = 0.8215 \]
Building the FRB with Chi-FRBCS-BigData-Ave

REDUCE

R_1: IF A_1 = L_1 AND A_2 = L_1 THEN C_1; RW_1 = 0.8743
R_2: IF A_1 = L_2 AND A_2 = L_2 THEN C_2; RW_2 = 0.9142
...

RB_1

R_1: IF A_1 = L_1 AND A_2 = L_1 THEN C_2; RW_3 = 0.9254
R_2: IF A_1 = L_1 AND A_2 = L_2 THEN C_2; RW_2 = 0.8842
...

RB_2

R_1: IF A_1 = L_2 AND A_2 = L_1 THEN C_2; RW_3 = 0.6534
R_2: IF A_1 = L_2 AND A_2 = L_2 THEN C_1; RW_1 = 0.7142
...

RB_3

R_1: IF A_1 = L_1 AND A_2 = L_1 THEN C_2; RW_1 = 0.2143
R_2: IF A_1 = L_3 AND A_2 = L_2 THEN C_2; RW_3 = 0.4715
...

RB_4

R_1: IF A_1 = L_2 AND A_2 = L_3 THEN C_2; RW_3 = 0.7784
R_2: IF A_1 = L_1 AND A_2 = L_1 THEN C_1; RW_2 = 0.8215
...

RB_n

Final RB generation

RB_1, R_1, C_1, RW = 0.8743
RB_2, R_1, C_2, RW = 0.9254
RB_3, R_2, C_1, RW = 0.7142
RB_4, R_1, C_2, RW = 0.2143
RB_5, R_2, C_1, RW = 0.8215

RC_1, C_1, RWave = 0.8033
RC_2, C_2, RWave = 0.5699
Summary of Chi-FRBCS-BigData

Homegenous fuzzy partitions shared by all Map tasks

Rule with equal antecedents are merged.

RWs are averaged

The original Chi fuzzy rule learning algorithm is applied

Consequent by P-CF

Cost-Sensitive Linguistic Fuzzy Rule Based Classification Systems under the MapReduce Framework for Imbalanced Big Data
V López, S del Río, JM Benítez, F Herrera
Experimental Analysis: Chi-FRBCS-BigData

- 6 Datasets with two classes problem
- Stratified 10 fold cross-validation
- Parameters:
  - Conjunction Operator: Product T-norm
  - Rule Weight: Penalized Certainty Factor
  - Number of fuzzy labels per variable: 3 labels

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Ex.</th>
<th>#Atts.</th>
<th>Selected classes</th>
<th>#Samples per class</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLCP</td>
<td>5749132</td>
<td>2</td>
<td>(FALSE; TRUE)</td>
<td>(5728201; 20931)</td>
</tr>
<tr>
<td>Kddcup_DOS_vs_normal</td>
<td>4856151</td>
<td>41</td>
<td>(DOS; normal)</td>
<td>(3883370; 972781)</td>
</tr>
<tr>
<td>Poker_0_vs_1</td>
<td>946799</td>
<td>10</td>
<td>(0; 1)</td>
<td>(513702; 433097)</td>
</tr>
<tr>
<td>Covtype_2_vs_1</td>
<td>495141</td>
<td>54</td>
<td>(2; 1)</td>
<td>(283301; 211840)</td>
</tr>
<tr>
<td>Census</td>
<td>141544</td>
<td>41</td>
<td>(-50000.; 50000+)</td>
<td>(133430; 8114)</td>
</tr>
<tr>
<td>Fars Fatal_Inj vs No_Inj</td>
<td>62123</td>
<td>29</td>
<td>(Fatal_Inj; No_Inj)</td>
<td>(42116; 20007)</td>
</tr>
</tbody>
</table>
Big Data Classification with Fuzzy Models

Analysis of the Performance, Precision

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Chi-FRBCS (\text{Acc}_{tr})</th>
<th>Chi-FRBCS (\text{Acc}_{tsl})</th>
<th>Chi-BigData-Max (\text{Acc}_{tr})</th>
<th>Chi-BigData-Max (\text{Acc}_{tsl})</th>
<th>Chi-BigData-Ave (\text{Acc}_{tr})</th>
<th>Chi-BigData-Ave (\text{Acc}_{tsl})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poker_0_vs_1</td>
<td>63.72</td>
<td>61.77</td>
<td>62.93</td>
<td>60.74</td>
<td>63.12</td>
<td>60.91</td>
</tr>
<tr>
<td>Covtype_2_vs_1</td>
<td>74.65</td>
<td>74.57</td>
<td>74.69</td>
<td>74.63</td>
<td>74.66</td>
<td>74.61</td>
</tr>
<tr>
<td>Census</td>
<td>96.52</td>
<td>86.06</td>
<td>97.12</td>
<td>93.89</td>
<td>97.12</td>
<td>93.86</td>
</tr>
<tr>
<td>Fars_Fatal_Inj_vs_No_Inj</td>
<td>99.66</td>
<td>89.26</td>
<td>97.01</td>
<td>95.07</td>
<td>97.18</td>
<td>95.25</td>
</tr>
<tr>
<td>Average</td>
<td>83.64</td>
<td>77.92</td>
<td>82.94</td>
<td>81.08</td>
<td>83.02</td>
<td>81.16</td>
</tr>
</tbody>
</table>

Good precision!
## Analysis of the Performance, Precision

<table>
<thead>
<tr>
<th>Datasets</th>
<th>16 mappers</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>32 mappers</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chi-BigData-Max</td>
<td>Chi-BigData-Ave</td>
<td>Chi-BigData-Max</td>
<td>Chi-BigData-Ave</td>
<td></td>
<td>Chi-BigData-Max</td>
<td>Chi-BigData-Ave</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc$_U$</td>
<td>Acc$_BT$</td>
<td>Acc$_U$</td>
<td>Acc$_BT$</td>
<td></td>
<td>Acc$_U$</td>
<td>Acc$_BT$</td>
<td>Acc$_U$</td>
<td>Acc$_BT$</td>
</tr>
<tr>
<td>Poker_0 vs 1</td>
<td>62.18</td>
<td>59.88</td>
<td>62.58</td>
<td>60.35</td>
<td></td>
<td>61.27</td>
<td>58.93</td>
<td>61.82</td>
<td>59.30</td>
</tr>
<tr>
<td>Covtype_2 vs 1</td>
<td>74.77</td>
<td>74.72</td>
<td>74.77</td>
<td>74.69</td>
<td></td>
<td>74.69</td>
<td>74.62</td>
<td>74.88</td>
<td>74.85</td>
</tr>
<tr>
<td>Census</td>
<td>97.44</td>
<td>93.75</td>
<td>97.45</td>
<td>93.52</td>
<td></td>
<td>97.11</td>
<td>93.48</td>
<td>97.12</td>
<td>93.32</td>
</tr>
<tr>
<td>Fars Fatal Ini vs No Ini</td>
<td>96.69</td>
<td>94.75</td>
<td>97.06</td>
<td>95.01</td>
<td></td>
<td>96.49</td>
<td>94.26</td>
<td>96.87</td>
<td>94.63</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>88.39</td>
<td>87.11</td>
<td>88.52</td>
<td>87.19</td>
<td></td>
<td>88.49</td>
<td>86.81</td>
<td>88.37</td>
<td>86.94</td>
</tr>
<tr>
<td></td>
<td>64 mappers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64 mappers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poker_0 vs 1</td>
<td>60.45</td>
<td>57.95</td>
<td>60.88</td>
<td>58.12</td>
<td></td>
<td>61.27</td>
<td>58.93</td>
<td>61.82</td>
<td>59.30</td>
</tr>
<tr>
<td>Covtype_2 vs 1</td>
<td>74.67</td>
<td>74.52</td>
<td>75.05</td>
<td>74.96</td>
<td></td>
<td>74.69</td>
<td>74.62</td>
<td>74.88</td>
<td>74.85</td>
</tr>
<tr>
<td>Census</td>
<td>97.07</td>
<td>93.30</td>
<td>97.13</td>
<td>93.41</td>
<td></td>
<td>97.11</td>
<td>93.48</td>
<td>97.12</td>
<td>93.32</td>
</tr>
<tr>
<td>Fars Fatal Ini vs No Ini</td>
<td>96.27</td>
<td>93.98</td>
<td>96.76</td>
<td>94.56</td>
<td></td>
<td>96.49</td>
<td>94.26</td>
<td>96.87</td>
<td>94.63</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>88.00</td>
<td>86.55</td>
<td>88.23</td>
<td>86.72</td>
<td></td>
<td>88.49</td>
<td>86.81</td>
<td>88.37</td>
<td>86.94</td>
</tr>
</tbody>
</table>

- **Performance improves slightly with less maps** (alleviate the small sample size problem)
- Chi-BigData-Ave obtains slightly better classification results
### Analysis of the Performance, Number of rules

<table>
<thead>
<tr>
<th>Kddcup_DOS_vs_normal dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NumRules by map</td>
<td>Final numRules</td>
</tr>
<tr>
<td>$RB_1$ size: 211</td>
<td>$RB_R$ size: 301</td>
</tr>
<tr>
<td>$RB_2$ size: 212</td>
<td></td>
</tr>
<tr>
<td>$RB_3$ size: 221</td>
<td></td>
</tr>
<tr>
<td>$RB_4$ size: 216</td>
<td></td>
</tr>
<tr>
<td>$RB_5$ size: 213</td>
<td></td>
</tr>
<tr>
<td>$RB_6$ size: 210</td>
<td></td>
</tr>
<tr>
<td>$RB_7$ size: 211</td>
<td></td>
</tr>
<tr>
<td>$RB_8$ size: 214</td>
<td></td>
</tr>
</tbody>
</table>

Robustness to the lack of data, increasing the final number of rules

**Table:**

<table>
<thead>
<tr>
<th>Class</th>
<th>Instance Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>972.781</td>
</tr>
<tr>
<td>DOS</td>
<td>3.883.370</td>
</tr>
</tbody>
</table>
### Big Data Classification with Fuzzy Models

**Analysis of the Performance, Number of rules**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Chi-FRBCS Average NumRules</th>
<th>Chi-BigData-Max Average NumRules</th>
<th>Chi-BigData-Ave Average NumRules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census</td>
<td>31518.3</td>
<td>34278.0</td>
<td>34278.0</td>
</tr>
<tr>
<td>Covtype_2_vs_1</td>
<td>6962.7</td>
<td>7079.1</td>
<td>7079.1</td>
</tr>
<tr>
<td>Fars_Fatal_Inj_vs_No_Inj</td>
<td>16843.3</td>
<td>17114.9</td>
<td>17114.9</td>
</tr>
<tr>
<td>Poker_0_vs_1</td>
<td>51265.4</td>
<td>52798.1</td>
<td>52798.1</td>
</tr>
</tbody>
</table>

**Robustness to the lack of data, increasing the final number of rules**

This may cause an improvement in the performance!!
## Big Data Classification with Fuzzy Models

### Analysis of the Performance, Runtime

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Chi-FRBCS Runtime (s)</th>
<th>Chi-BigData-Max Runtime (s)</th>
<th>Chi-BigData-Ave Runtime (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census</td>
<td>38655.60</td>
<td>1102.45</td>
<td>1343.92</td>
</tr>
<tr>
<td>Covtype_2_vs_1</td>
<td>86247.70</td>
<td>2482.09</td>
<td>2512.16</td>
</tr>
<tr>
<td>Fars_Fatal_Inj_vs_No_Inj</td>
<td>8056.60</td>
<td>241.96</td>
<td>311.95</td>
</tr>
<tr>
<td>Poker_0_vs_1</td>
<td>114355.80</td>
<td>5672.80</td>
<td>7682.19</td>
</tr>
<tr>
<td>Average</td>
<td>61828.93</td>
<td>2374.82</td>
<td>2962.56</td>
</tr>
</tbody>
</table>

### KddCUP’99

<table>
<thead>
<tr>
<th>Class</th>
<th>Instance Number</th>
<th>Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>972.781</td>
<td>8</td>
</tr>
<tr>
<td>DOS</td>
<td>3.883.370</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>132</td>
</tr>
</tbody>
</table>
Fuzzy-based models for Big Data learning

OUTLINE

- Fuzzy Rule Based Systems
- Data fragmentation and lack of data problem
- Big Data classification with fuzzy models
- Are FRBCSs robust with respect to the lack of data?
- Conclusions and future challenges
Experimental Framework

Two versions of Chi algorithm will be used:

- Chi-FRBCS-BigData (noted as ChiBD): MapReduce implementation
- Sequential Chi (noted as ChiStd): Results of 1Map (from 1 to 128)

Table 1. Summary of BigData classification problems.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Ex.</th>
<th>#Atts.</th>
<th>Selected classes</th>
<th>#Samples per class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covtype_1_vs_2</td>
<td>495,173</td>
<td>54</td>
<td>(1; 2)</td>
<td>(211,705; 283,468)</td>
</tr>
<tr>
<td>Poker_0</td>
<td>1,025,009</td>
<td>10</td>
<td>(0; remainder)</td>
<td>(513,701; 511,308)</td>
</tr>
<tr>
<td>Susy</td>
<td>4,923,622</td>
<td>18</td>
<td>(0;1)</td>
<td>(2,711,811; 2,211,811)</td>
</tr>
</tbody>
</table>

Table 2. Configuration parameters for Chi-FRBCS-BigData.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Labels:</td>
<td>3 fuzzy partitions</td>
</tr>
<tr>
<td>Conjunction operator:</td>
<td>Product T-norm</td>
</tr>
<tr>
<td>Rule Weight:</td>
<td>Penalized Certainty Factor</td>
</tr>
<tr>
<td>Fuzzy Reasoning Method:</td>
<td>Winning Rule</td>
</tr>
</tbody>
</table>
Are FRBCS robust with respect to the lack of data?

Good precision

Accuracy obtained by Chi-FRBCSBigData (ChiBD) and a sequential run of Chi (ChiStd) with respect to Maps
Analysis for the number of rules within Maps

- Higher #maps, the smaller the contribution of each RBi.
- Better interpretability within each local fuzzy classifier,
- Few fuzzy rules: good representation of the problem space
- Global RB composed by more rules (RBi aggregation)
Analysis for the number of rules within Maps

Total Rules obtained by **ChiBD** with respect to #Maps – Covtype1vs2

<table>
<thead>
<tr>
<th>Maps</th>
<th>Total Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6100</td>
</tr>
<tr>
<td>8</td>
<td>6150</td>
</tr>
<tr>
<td>16</td>
<td>6200</td>
</tr>
<tr>
<td>32</td>
<td>6250</td>
</tr>
<tr>
<td>64</td>
<td>6300</td>
</tr>
<tr>
<td>128</td>
<td>6350</td>
</tr>
</tbody>
</table>

% Filtered Rules in **ChiBD** with respect to #Maps – Covtype1vs2

<table>
<thead>
<tr>
<th>Maps</th>
<th>Filtered Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>23.00%</td>
</tr>
<tr>
<td>16</td>
<td>24.00%</td>
</tr>
<tr>
<td>32</td>
<td>25.00%</td>
</tr>
<tr>
<td>64</td>
<td>26.00%</td>
</tr>
<tr>
<td>128</td>
<td>27.00%</td>
</tr>
</tbody>
</table>

**More Maps → More Rules.**

- P-CF RW: **Filter** rules (RW < 0) from local RBs.
- Sparsely distributed data: values of RWs become higher.
Fuzzy rules in Reduce: Same antecedent

- High number of co-occurrences, especially for a low # maps:
  - The larger the volume of input data, more similar clusters are expected.
  - Same fuzzy labels chosen from these clusters
- Equal rules found among maps: higher density of data within the problem space.
Fuzzy rules in Reduce: Double consequent

- How many rules in conflict arrive to a Reduce task.
- Rules related to overlapped regions.
- Low percentage:
  - Probably found within a single map process, and discarded due to the RW computation
Fuzzy rules in Reducer: Unique rules

- Rules generated in a single Map task.
- This value is practically the same for all Map case studies.
- Instances far from areas of high density of data.
- Their influence is independent of the data distribution.
Fuzzy-based models for Big Data learning

OUTLINE

- Fuzzy Rule Based Systems
- Data fragmentation and lack of data problem
- Big Data classification with fuzzy models
- Are FRBCSs robust with respect to the lack of data?
- Conclusions and future challenges
Concluding Remarks

- **Linguistic fuzzy models for Big Data** under the MapReduce framework:
  - Manages big datasets
  - Without damaging the classification accuracy
  - Fast response times (increasing with the number of Maps)
- **Robustness** of fuzzy models when addressing the problem of **small disjuncts** and **lack of data**:
  - High degree of *replicated* rules along maps: partial RBs represent accurately a high percentage of the problem.
  - *Local models* learned in each Map task are of *high quality*. Aggregation boosts the recognition ability of the FRBCS.
  - Those fuzzy rules that are obtained from more than a map *contribute the most* to the final classification
Future Challenges

- Designing new learning methodologies for FRBCS in MapReduce:
  - Reduce phase for approximate fuzzy models
  - Deep analysis “ensembles vs. fusion of rules”
  - Analysis on small disjuncts preprocessing for fuzzy models
  - New fuzzy models based on accurate algorithms
  - Study the effect of RW computation in the quality of the fuzzy model
  - Application of good practices by means of scalable models:
    - Contextualization of the Data Base
    - Optimization of the Rule Base

- A promising line of work for the design of high performance Fuzzy Models for Big Data
Big Data Classification with Fuzzy Models

Code for our approaches: https://github.com/aFdezHilario

Alberto Fernandez Hilario

Fuzzy Rule Based System for classification (w/wo cost sensitive)

Evolutionary Fuzzy System for Rule Selection
Outline

- Introduction to Big Data. Big data analytics
- Fuzzy-based models for Big Data Learning
- Smart Data: The missing bridge between Big Data and real applications to get high quality data
- Fuzzy Big Data Science: Opportunities
- Final Comments
“Without Analytics, Big Data is just Noise”

Guest post by Eric Schwartzman, founder and CEO of Comply Socially

http://www.briansolis.com/2013/04/without-analytics-big-data-is-just-noise/

Hal Varian
Google Chief Economist

The challenge is to translate raw data into information that can be used to improve performance.
What are the challenges of analyzing Big Data?

Big data are characterized by high dimensionality and large sample size. These two features raise three unique challenges: (i) **high dimensionality brings noise accumulation, spurious correlations** and incidental homogeneity, (ii) high dimensionality combined with large sample size creates heavy computational cost and algorithmic instability; (iii) the massive samples in Big Data are typically aggregated from multiple sources at different time points using different technologies. This creates issues of heterogeneity, experimental variations and statistical biases, and requires us to develop more adaptive and robust procedures.
RECALL: Big data as a concept is defined around five aspects:

- Data volume,
- Data velocity,
- Data variety,
- Data veracity and
- Data value.
a) While the *volume, variety and velocity* aspects refer to data generation process and how to capture and store the data,

b) *Veracity and value* aspects deal with the *quality and the usefulness of the data* leading to the point.
Smart Data

Smart Data (veracity and value) aims to filter out the noise and hold the valuable data, which can be effectively used by enterprises and governments for planning, operation, monitoring, control, and intelligent decision making.

What makes data smart? Three key attributes for data to be smart, it must be **accurate**, **actionable**, and **agile**.
Big Data (Analytics) → Smart Data

The key is to explore how Big Data can become Smart Data.

“Without Analytics, Big Data is just Noise”

*Eric Schwartzman*

Big Data + Analytics = Smart Data

“Here’s a list of 100,000 warehouses full of data. I’d like you to condense them down to one meaningful warehouse.”
Big Data (Analytics) → Smart Data

Smart Data

Quality data for quality decisions!

Big Data

- Data Science
  - Data Preprocessing
  - Predictive and descriptive Analytics
  - Model building
Big Data (Analytics) \( \rightarrow \) Smart Data

What is included in data preprocessing?

![Diagram of data preprocessing steps]

Fig. 1.3 Forms of data preparation

Fig. 1.4 Forms of data reduction

S. García, J. Luengo, F. Herrera
Data Preprocessing in Data Mining, Springer, 2015
Big Data Preprocessing → Smart Data

Big data preprocessing also must spend a very important part of the total time in a big data analytic process.

Credit: http://www.ibm systemsmag.com/aix/storage/servers/The-Data-Squeeze/
Big Data: Algorithms for Data Preprocessing, Computational Intelligence, and Imbalanced Classes

The web is organized according to the following summary:

1. Introduction to Big Data
2. Big Data Technologies: Hadoop ecosystem and Spark
3. Big Data preprocessing
4. Imbalanced Big Data classification
5. Big Data classification with fuzzy models
6. Classification Algorithms: k-NN
7. Big Data Applications
8. Dataset Repository
9. Literature review: surveys and overviews
10. Keynote slides
11. Links of interest
Feature Selection

Fast-mRMR: an optimal implementation of minimum Redundancy Maximum Relevance algorithm

https://github.com/sramirez/fast-mRMR

This is big data based implementation of the classical feature selection method: minimum Redundancy and Maximum Relevance (mRMR); (Hanchuan Peng, Fuhui Long, and Chris Ding "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005).

This includes several optimizations such as: cache marginal probabilities, accumulation of redundancy (greedy approach) and a data-access by columns.
Feature Selection

**Fast-mRMR:** Minimum Redundancy Maximum Relevance algorithm

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#Train Ex.</th>
<th>#Test Ex.</th>
<th>#Atts.</th>
<th>#Cl.</th>
<th>Sparse</th>
</tr>
</thead>
<tbody>
<tr>
<td>epsilon</td>
<td>400 000</td>
<td>100 000</td>
<td>2000</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>dna</td>
<td>79 739 293</td>
<td>10 000</td>
<td>200</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>ECBDL14</td>
<td>65 003 913</td>
<td>2 897 917</td>
<td>630</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>url</td>
<td>1 916 904</td>
<td>479 226</td>
<td>3 231 961</td>
<td>2</td>
<td>Yes</td>
</tr>
<tr>
<td>kddb</td>
<td>19 264 097</td>
<td>748 401</td>
<td>29 890 095</td>
<td>2</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**TABLE IV:** Selection time by dataset and threshold (in seconds)

<table>
<thead>
<tr>
<th># Features</th>
<th>kddb</th>
<th>url</th>
<th>dna</th>
<th>ECBDL14</th>
<th>epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>283.61</td>
<td>94.06</td>
<td>97.83</td>
<td>332.90</td>
<td>111.42</td>
</tr>
<tr>
<td>25</td>
<td>774.43</td>
<td>186.22</td>
<td>148.78</td>
<td>596.31</td>
<td>173.39</td>
</tr>
<tr>
<td>50</td>
<td>1365.82</td>
<td>333.70</td>
<td>411.84</td>
<td>1084.58</td>
<td>292.07</td>
</tr>
<tr>
<td>100</td>
<td>2789.55</td>
<td>660.48</td>
<td>828.35</td>
<td>2420.94</td>
<td>542.05</td>
</tr>
</tbody>
</table>

*Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data.*
From high dimensionality to small dimensionality:

Eases the application of fuzzy learning systems

**TABLE IV:** Selection time by dataset and threshold (in seconds)

<table>
<thead>
<tr>
<th># Features</th>
<th>kddb</th>
<th>url</th>
<th>dna</th>
<th>ECBDL14</th>
<th>epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>283.61</td>
<td>94.06</td>
<td>97.83</td>
<td>332.90</td>
<td>111.42</td>
</tr>
<tr>
<td>25</td>
<td>774.43</td>
<td>186.22</td>
<td>148.78</td>
<td>596.31</td>
<td>173.39</td>
</tr>
<tr>
<td>50</td>
<td>1365.82</td>
<td>333.70</td>
<td>411.84</td>
<td>1084.58</td>
<td>292.07</td>
</tr>
<tr>
<td>100</td>
<td>2789.55</td>
<td>660.48</td>
<td>828.35</td>
<td>2420.94</td>
<td>542.05</td>
</tr>
</tbody>
</table>

**Data Set** | **#Atts.**
------------|--------------
epsilon     | 2000         
dna         | 200          
ECBDL14     | 630          
url         | 3 231 961    
kddb        | 29 890 095   

**Fast-mRMR:** Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data.
Outline

- Introduction to Big Data. Big data analytics
- Fuzzy-based models for Big Data Learning
- Smart Data: The missing bridge between Big Data and real applications to get high quality data
- Fuzzy Big Data Science: Opportunities
- Final Comments
Fuzzy Big Data Science: Opportunities

There are open multiple scenarios of data science and their challenges and opportunities for fuzzy modeling:

- Multi Label
Fuzzy Big Data Science: Opportunities

There are open multiple scenarios of data science and their challenges and opportunities for fuzzy modeling:

- Multi Label
- Multi Instance Learning
There are open multiple scenarios of data science and their challenges and opportunities for fuzzy modeling:

- Multi Label
- Multi Instance Learning
- Multiview Learning
Fuzzy Big Data Science: Opportunities

There are open multiple scenarios of data science and their challenges and opportunities for fuzzy modeling:

- Multi Label
- Multi Instance Learning
- Multi View Learning
- Semi-supervised Learning
- Transfer Learning
- Big Data
“Going beyond straightforward fuzzy extensions of conventional ML methods, we need to focus on the right topics, correctly appraise the (complementary) role of fuzzy sets in learning from data, and avoid unwarranted claims”.
Outline

- Introduction to Big Data. Big data analytics
- Fuzzy-based models for Big Data Learning
- Smart Data: The missing bridge between Big Data and real applications to get high quality data
- Fuzzy Data Science: Beyond Interpretability-Accuracy Tradeoff
- Fuzzy Big Data Science: Opportunities
- Final Comments
Final Comments

Data Mining, Machine learning and data preprocessing: Huge collection of algorithms

Big Data Analytics

Big Data: A small subset of algorithms

Big Data Preprocessing:
A few methods for preprocessing in Big Data analytics. Evolutionary models are a promising approach.

Soft computing approaches are useful to tackle big data analytics
Final Comments

- **Fuzzy models for big data:** Robustness to the lack of data for the data fragmentation, producing high performance (accuracy).

- **The focus should be on**
  - a) designing models for maps and
  - b) the combination phase (reduce).

- **Potential applications in multiple scenarios:**

<table>
<thead>
<tr>
<th>BigData&amp; Analytics and Non Standard Classification</th>
<th>Social Big Data</th>
</tr>
</thead>
</table>

93
Final Comments: Questions for discussion

- **EH**: “Yet, interesting contributions to fuzzy machine learning have already been made, and even more significant ones are conceivable.”

- **Interpretability, complexity, interpretation.** Are we caught in a spiral? Why more discussions on interpretability? What are the limits?

- **Beyond classical classification and regressions.** What are the possibilities of fuzzy systems in the new data science areas?

- **Big Data.** Why to design fuzzy systems for big data? Real scenarios.

- **Deep learning in the center of the Artificial Intelligence. Deep Learning vs fuzzy systems.** Why to use fuzzy systems? What is the position of Fuzzy logic in the Artificial Intelligence board?

  N. Pal: “Designing Intelligent Fuzzy Systems performing well, understandable and simple”

  “We have to place the Fuzzy Systems on the stage where they can play an important role in Data Science”.
Final Comments: Questions for discussion

We must explain who to use fuzzy systems with cases of study in 20 seconds

Example with Deep learning: Handgun alarm detection in videos

We have a robbery in a jewelry store, and the thief carries a gun and it is detected in 1 second giving an alarm
Quality decisions must be based on quality big data and quality models. Computational Intelligence may be very useful to create new high quality big data analytics models.
Thanks!!!