Fuzzy Based Decision Quality Evaluation based on Measures of Interestingness

Shinde-Pawar Manisha, Jamsandekar Pallavi

Abstract: Social Media and sentiment analysis is a glamorous field of attraction for researchers. Amplified data generation and its challenges need significant solutions to satisfy the needs of investigation or to discover relevant and required glimpse to support decision in domain of study. Opinion mining is knowing attitude if author of content tracking general attitude about the topic of content or product or service. The study reveals the systematic framework for analysis of quality of selection and application of measures of interestingness. This framework involves fuzzy based method for classification, prioritization of decisions against measures to define and validate the novel method and results in data mining. The researcher aims to value work of fiction of evaluation in data mining.

Index Terms: Data, Decision, Fuzzy based method, Measures of Interestingness, Opinion mining, Sentiment Analysis, Social Media.

1. INTRODUCTION

Measures of interestingness play vital role in different stages of data mining and sentiment analysis. Internet based social media data, data warehouses, web applications, and blog data, data base all these sources represents and holds data in different forms.

2. Data to decision is stage wise systematic process goes through different parameters of evaluation and verification of stage output through different measures as shown below in figure no.1:

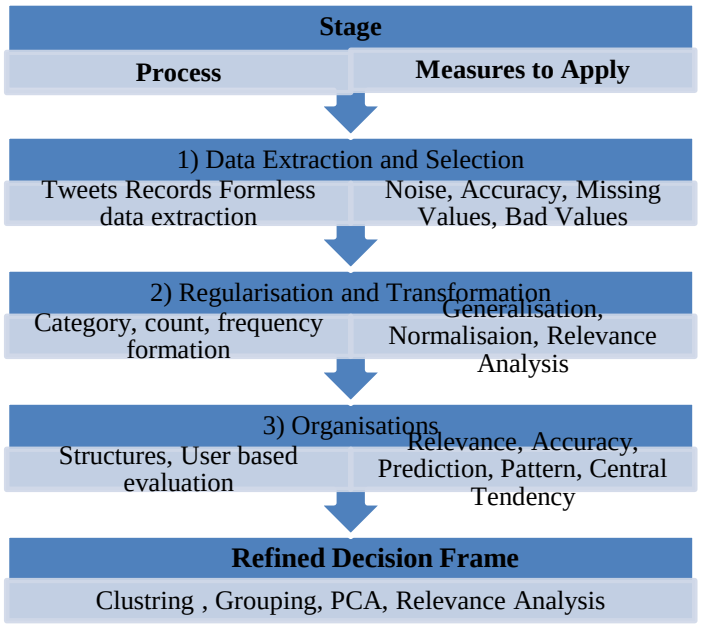

Fig. 1 Data Analysis

2.1 DATA EXTRACTION AND SELECTION:

Facebook, Twitter, WhatsApp data is formless data. So right from initial stage of data extraction, challenges and obstacles need to be passed to reach to target and discover relevant valuable knowledge. As shown in figure no.1, vey first stage of data extraction needs to deal with noise, accuracy, missing values and bad value measure parameters mainly. By verifying with these measures, Data Quality for completeness, correctness, accuracy, consistency and integrity is ensured.

- *Shinde-Pawar Manisha, Jamsandekar Pallavi*
- *Assistant Professor, [2]Professor & HOD*
- *Bharati Vidyapeeth (Deemed to be University), Pune*
- *Institute of Management & Rural Development Administration, Sangli mjs.imrda@gmail.com*
- *pallavi.jamsandekar@yahoo.com*

2.2 REGULARIZATION AND TRANSFORMATION:

Data confirmed for its quality is further regularized and normalized to bring it to general form from specific. Standardization and regularization brings ill-defined data in order or simple form and so performs selection technique to control the complexity of data.

The measures as generalization, normalization with specified rule sets or task set and relevance analysis are performed to verify data transformed to consistent, normal and decomposed to simple form.

2.3 ORGANIZATION AND STRUCTURING:

User prefers according to data domain, data source and data form may be applied to restructure data or to compose organized form of data to reach to formed from formless so data with proper relevance and other measures will be arranged in organized manner of its association. User based measures and reduction methods can be applied to get real sample representative of data. For Example: Some clustering, grouping, principal component analysis, factor analysis or relevance analysis. It involves reducing complexity and reaching to more simple/ normal form without actually missing meaning or real glimpse of information in data.

3 DECISION DISCOVERY:

In this last phase, different algorithms of methods are implemented to estimate values, to get certainty, to classify data, to identify patterns or to predict data based on available data for the given data domain. Objective measures based on statistics and structure of patterns (counts, frequency, probability distribution function). And subjective measures are based on user belief about data. These measures can be applied after and before the implementation of logical approaches to reach to significant decision. More significantly for different situation and process, method and data type demand of measures combination is different. Generally with literature survey or using Delphi technique measures are selected and applied to process, so resultant decision quality

may be based on results of measures of every stage.

Use of mix of refined statistical and high performance computing techniques can lead in better evaluation techniques for performance analysis. This mix decision in light of number of subjective and objective measures availability becomes vague for so many dynamic conditions and constraints. Fuzzy logic can address the vagueness and complexity to classify, prioritize the measures according to different dynamic input constraint values. As stage wise measure results can be assigned to input membership parameters according to result of measures. The combination of these multiple parameters with different membership interval may lead in different certain value can be analyzed as evaluation measure of decision quality.

3.1 FUZZY APPROACH FOR MEASURES OF INTERESTINGNESS:

R TOOLS USED FOR ANALYSIS:

Researcher has applied the experiment using R- programming tool.

- R Programming
- RStudio Version- 1.1.383

R Programming provides many packages for data analysis and data mining.

3.2 RULE ANALYSIS:

As result of measures may be different for different stages of data processing in sentiment analysis. Mix of premises in inferences may change the value of resultant class in fuzzy rules. Fuzzy rules in its inference may be used with connection for multiple operations with AND, OR, NOT. In the proposed fuzzy system, fuzzy input parameters and variables partitions are made for values sub ranges and again connected using AND, OR connectors as shown in table no.1.

Table 1

Measures Result Parameter Rule Base

DataQuality	DataTransform	DataStructure	Output Variable DecisionQuality
VeryHigh	VeryHigh	VeryHigh	VeryStrong
High	High	High	Strong
Low	Low	Low	Weak
VeryLow	VeryLow	VeryLow	VeryWeak

Source: Compiled by Researcher

Different three stages (DataQuality, DataTransform, DataStrcture) and their measure result can be passed as parameters ranging to different values as Very High, High, Low, VeryLow which determines resultant output class variable DecisionQuality as one of the partitions specified VeryWeak, Weak, Strong or VeryStrong.

a drop cap.

4 INPUT PARAMETERS:

Decision quality of automated analytics is determined by using different determinants and reliability and extent of role played by theses determinants in decision making will be different every time. Different levels of values of input membership are captured using partitions framed for three input variable are as:

4.1 DATAQUALITY:

Table 2

DataQuality Input Variable Fuzzy Parameter Partitions

DataQuality Label	Class	Degree of Class	Meaning
VeryLow	NotReliable	20	Data Quality is very less and not reliable need more cleansing and pre-processing
Low	LessReilable	40	Data Quality is low may need cleansing and pre-processing
High	Reliable	60	Data Quality is high and reliable
VeryHigh	HighlyReliable	80	Data Quality is Very High more reliable

Source: Compiled by Researcher

In very first stage, data is pre-processed to deal with missing value and noisy data to bring it to more reliable form as depicted in table No.2. This pre-processing is performed to get data quality improved but methods and practices used for pre-processing and original data validity results in different extents of data quality, to capture granularity of such data preprocessed data quality interestingness measure applied will result value which can be recognized with different sub-ranges of partitions for input membership fitness.

4.2 DATATRANSFORM:

Table 3

DataTransform Input Variable Fuzzy Parameter Partitions

DataTransform Label	Class	Degree of Class	Meaning
VeryLow	Formless	20	Data is not transformed to generalized or regularized form
Low	Less-Formed	40	Data is transformed but less regularized
High	Formed	60	Data is transformed and regularized at normal level
VeryHigh	Most-Formed	80	Data is well transformed and regularized at high level

Source: Compiled by Researcher

Data Transformation is second stage of sentiment analysis where using count of terms, frequency and / or probability normal distribution pre-processed data is generalized or regularized to bring it to well form from formless. The resultant generalization and level of form is passed as input range of DataTransform member to different degree of classes as shown in Table no. 3.

4.3 DATASTRUCTURE:

Table 4

DataStructure Input Variable Fuzzy Parameter Partitions

Data Structure Label	Class	Degree of Class	Meaning
VeryLow	Non Structured	20	Data is Not structured and not relevant
Low	Structured	40	Data is less structured and less relevant
High	Highly Structured	60	Data is structured and relevant
VeryHigh	Very Highly Structured	80	Data is highly structured and highly relevant

Source: Compiled by Researcher

Table No. 4 depicts the partition of DataStructure partitions to get ranges of structure of data. Data well structured is always more relevant and accurate. So relevance analysis, pattern detection, accuracy and central tendency or any user based measure evaluates data structures formed.

5 OUTPUT VARIABLE: DECISIONQUALITY:

Decision is discovered in last stage in sentiment analysis as value decision as data is stage wise processed to reach to its all requirement satisfaction for qualitative decisions. But quality may vary depending on quality of each method and process application.

Table 5
Fuzzified DecisionQuality Classification Output Partitions

DecisionQuality Label	Class	Degree of Class	Meaning
VeryWeak	NotQuality Decision	20	Very less decision quality
Weak	LessQuality Decision	40	Less Decision Quality
Strong	StrongQuality Decision	60	Strong decision quality
VeryStrong	VeryStrongQuality Decision	80	Very Strong decision quality

Source: Compiled by Researcher

Finally combination of membership partition of all input parameters for measures evaluated leading in final decision will decide whether the resultant DecisionQuality is very weak, weak, strong or very strong as represented in Table No.5

6. FUZZY IMPLEMENTATION

By using fuzzy system model shown in figure no.2, 8 rules for fuzzy system with three input and one output variable that 4 variables fuzzy system is constructed.

>print(system)
A fuzzy system consists of 4 variables and 8 rules.

Variables:

DataTransform(VeryLow, Low, High, VeryHigh)
DecisionQuality(VeryWeak, Weak, Strong, VeryStrong)
DataStructure(VeryLow, Low, High, VeryHigh)
DataQuality(VeryLow, Low, High, VeryHigh)

Rules:

1. DataQuality %is% Low || DataQuality %is% VeryLow && DataTransform %is% VeryLow || DataTransform %is% Low && DataStructure %is% High => DecisionQuality %is% VeryWeak
2. DataQuality %is% Low || DataQuality %is% VeryLow && DataTransform %is% VeryLow || DataTransform %is% Low && DataStructure %is% VeryHigh => DecisionQuality %is% VeryWeak
3. DataQuality %is% Low || DataQuality %is% VeryLow && DataTransform %is% VeryLow || DataTransform %is% Low && DataStructure %is% Low => DecisionQuality %is% Weak
4. DataQuality %is% Low || DataQuality %is% VeryLow && DataTransform %is% VeryLow || DataTransform %is% Low && DataStructure %is% VeryLow => DecisionQuality %is% Weak
5. DataQuality %is% VeryHigh || DataQuality %is% High && DataTransform %is% VeryHigh || DataTransform %is% High && DataStructure %is% Low => DecisionQuality %is% Strong
6. DataQuality %is% VeryHigh || DataQuality %is% High && DataTransform %is% VeryHigh || DataTransform %is% High && DataStructure %is% VeryLow => DecisionQuality %is% Weak
7. DataQuality %is% VeryHigh || DataQuality %is% High && DataTransform %is% VeryHigh || DataTransform %is% High && DataStructure %is% VeryHigh => DecisionQuality %is% VeryStrong
8. DataQuality %is% VeryHigh || DataQuality %is% High && DataTransform %is% VeryHigh || DataTransform %is% High && DataStructure %is% High => DecisionQuality %is% VeryStrong

> plot(system) ## plots variables

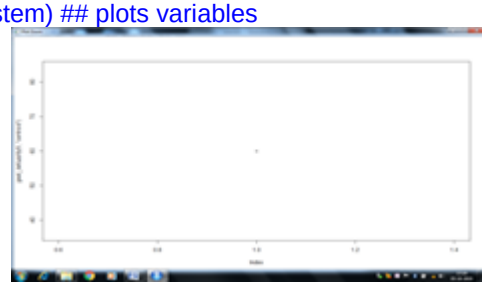

Fig. 2 *Fuzzy System*

The fuzzy system with different classes of input and output membership inferring to 8 rules set is evaluated for input value combination as below:

> fi <- fuzzy_inference(system, list(DataQuality = 80, DataTransform = 75, DataStructure = 60))
> gset_defuzzify(fi, "centroid")
[1] 60
> plot(gset_defuzzify(fi, "centroid")) #

As result of inference input range DataQuality= 80, DataTransform = 75 and DataStructure = 60 gives resultant values fuzzified gets defuzzified to 60 using centroid method of defuzzification as shown in figure no. 3.

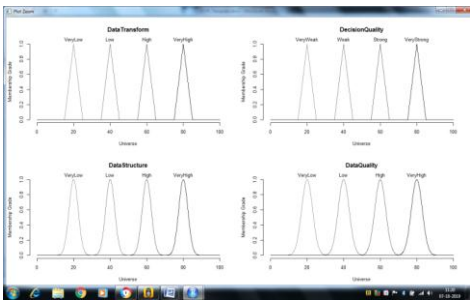

Fig. 3 : *Defuzzified value Centroid for certain decision.*

4 CONCLUSION

The research work considers fuzzy based approach for decision generated using systematic framework; each phase is evaluated with different measures of interestingness. Measures result and evaluation aspect can be significantly developed to focus different dimensions of evaluation at each phase. In future research, the researcher has planned to determine approach of dynamic selection of measures of interestingness. Further the researcher would like to apply classification to identify suitable measures as pruning method to reach to considerable rules from number of rules available and results combination for sentiment analysis. By applying algorithms to draw patterns of measures of interestingness and to identify members fitting to particular pattern are general processes wherein algorithm and pattern creation methods may different in different cases.

REFERENCES

[1] Ashish Shukla, Rahul Misra, "Sentiment Classification and Analysis Using Modified K-Means and Naïve Bayes Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, Issue 8, Aug. 2015

[2] BEN, "On Social Sentiment And Sentiment Analysis", DECEMBER 16, 2013 , http://brnrd.me/social-sentiment-sentiment-analysis/

[3] Hemant Sharma, "What is Data Scienec? A Beginner's Guide to data science", .https://www.edureka.co/blog/what-is-data-science/ published on 24th Dec 2018.

[4] J. I. Sheeba and Dr. K. Vivekanandan[2014], "A Fuzzy Based Sentiment Classification",International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.4, pp no(27-44)

[5] Ken Mcgarry (2005), "A Survey of Interestingness Measures for Knowledge Discovery", The Knowledge Engineering Review, Vol. 00:0, 1–24, Cambridge University Press

[6] Maria Sokolova, Guy Lapalme(2009), "A Systematic analysis of performance measures for classification tasks.", Information Processing and Management 45 (2009) 427–437

[7] Mojdeh Jalali-Heravi, Osmar R. Zaïane, "A Study on Interestingness Measures for Associative Classifiers", University of Alberta, Canada, 2010Steven Finlay, "Predictive Analytics, Data Mining and Big Data", published in 2014 by Palgrave MacMillan.

[8] Timothy Dauria(2012)," How to Build a Text Mining, Machine Learning Document Classification System in R!", Published on May 16, 2012 on Youtube

[9] https://www.tidytextmining.com/sentiment.html

[10] https://en.wikipedia.org/wiki/Data_science