Examiner's report of doctoral thesis

Author: Ing. Raphael Kwaku Botchway

Title: Soft Computing Techniques for Sentiment Analysis and Feature Selection

Examiner: prof. RNDr. PaedDr. Eva Volná, PhD.
University of Ostrava

Objectives of the thesis and their fulfilment

A given doctoral thesis is focussed on soft computing techniques for sentiment analysis. This thesis was structured into two segments. The first segment utilizes the abundance of social media data available online to explore the use of soft computing techniques for sentiment analysis. The second part use of evolutionary computation techniques to solve feature selection problems. In this phase, a metaheuristic-based solution using the Particle Swarm Optimization (PSO) algorithm for optimal subset text feature selection during sentiment analysis was implemented.

Chosen methods can be considered as fully competent and it is sufficient for purposes and objectives of the PhD thesis. The given thesis has fulfilled its main objectives and its topic is up to date.

Benefits in the field of knowledge

Author proposed and implemented a new metaheuristic-based solution using the Particle Swarm Optimization algorithm and its hybridization for optimal subset text feature selection in the task of sentiment analysis. Author used a low-level coevolutionary mixed hybrid approach to develop a new hybrid metaheuristic algorithm AMGWOPSO by hybridizing the GWO with the AMPSO for wrapper feature selection. The resulting hybrid metaheuristic algorithm has embraced the angle modulation technique used in the domain of signal processing within the telecommunication industry. An ability to use the AMGWOPSO as a wrapper feature selection method for feature selection constitutes a major contribution to this work.

The proposed procedures are supported by several publications of the doctoral student. In the years 2019-2023, he published 4 articles in journals (3 of them with IF) and 7 papers in conference proceedings.

Přírodovědecká fakulta Ostravské univerzity / 30. dubna 22 701 03 Ostrava / Česká republika
e-mail: lucie.zarnikova@osu.cz / telefon: +420 597 092 102
prf.osu.cz / alive.osu.cz

Benefits in the field of social practice

The results of sentiment analysis in this work demonstrate that firms need to focus on customer engagement to enhance customer experience via social media channels. Furthermore, a data-driven approach produced robust outputs compared to conventional marketing approaches.

It is obvious that the presented proposals are based on good author's knowledge and experience with an implementation of similar problems in practice.

Formal arrangement

The doctoral thesis has 99 pages. The whole thesis is written in English. The thesis is written and structured in a logical and well-arranged way. Its text is presented at an appropriate level of expertise and it is compact, only some of the images are not of adequate quality.

Questions and comments

1. In the theoretical part you mentioned several evolutionary techniques (PSO, grey wolf optimization, whale optimization algorithm, AMPSO), but this part has no summary. Could you compare these algorithms and explain what you used from the afore mentioned to design your algorithm?

2. Could you propose a continuation of this work?

Conclusion

The submitted thesis fulfils the requirements for a doctoral thesis, both in terms of theoretical - methodological level, so the usefulness in practice. The thesis contains the original results.

I recommend the thesis to the defence before the relevant commission. Based on the thesis, I suggest the academic and scientific degree "Doctor Philosophiae" (Ph.D. abbreviation) to confer to Ing. Raphael Kwaku Botchway after successfully defending of his thesis.

Ostrava, 1 September 2023   Prof. RNDr. PaedDr. Eva Volná, PhD.

Přírodovědecká fakulta Ostravské univerzity / 30. dubna 22 701 03 Ostrava / Česká republika
e-mail: lucie.zarnikova@osu.cz / telefon: +420 597 092 102
prf.osu.cz / alive.osu.cz

Oponentský posudek dizertační práce Ing. Raphaela Kwaku Botchwaye "Soft Computing Techniques for Sentiment Analysis and Feature Selection"

a) Aktuálnost zvoleného tématu

Práce je velmi aktuální, analýza sentimentu spolu s výběrem příznaků se dostává do popředí v posledním desetiletí s neustálým zvyšováním nutnosti automaticky analyzovat sentiment na www, až už jde o zákazníky obchodu nebo politickou kampaň. Zároveň jsou možnosti těchto analýz podporovány stále vyšším výkonem počítačů. Námět odpovídá oboru disertace.

b) Splnění sledovaných cílů

Cíle stanovené v práci na str. 33 byly jednoznačně splněné.

c) Zvolené metody zpracovaní

Celá práce je psaná přehledně na výborné anglické jazykové úrovni, prakticky bez překlepů (v druhém řádku tabulky 6.5 na straně 47 myslím chybí první číslice u prvního čísla). Zvolené přístupy jsou moderní, kombinace různých optimalizačních přístupů dává často nejlepší výsledky.

d) Dosažené výsledky / nové poznatky

Výsledky práce přinášejí nové poznatky v oblasti analýzy sentimentu a výběru příznaků.

Publikační aktivita doktoranda je kvalitní, kromě jiného je jako první autor dvou článků v hodnotných časopisech (oba Q2 kvartil), příspěvku na IEEE konferencí a v ACM sborníku, v dalších časopisech a konferenčních sbornících je jako další ze spoluautorů, což ukazuje jeho schopnost pracovat v kolektivu.

e) Přínos pro další rozvoj vědy

Přínosy pro další rozvoj vědy a techniky jsou jednoznačné, kvalitativně vyjádřené zpřesněním a zrychlením metody výběru příznaků. Originální přínos je především ve vytvoření hybridní techniky evolučních výpočtů s úhlovou modulací pro řešení problému výběru příznaků. Kvalitu modifikace zpracování sentimentu nejsem až tak dobře schopen posoudit a kvalita výběru použití binární PSO spolu s SVM při analýze sentimentu byla například potvrzena cizí prací (Yadav, A. and Vishwakarma, D.K., 2020. A comparative study on bio-inspired algorithms for sentiment analysis. Cluster Computing, 23, pp.2969-2989). Dizertant má vědeckou erudici, prokázal schopnost samostatně a tvořivě řešit vědecké problémy dané oblasti.

Na dizertanta mám následující otázky:

1. Na porovnání algoritmů pro selekci příznaků jste využil Python kódu z githubu, přičemž Vámi navrhnutý algoritmus s úhlovou modulací vyšel jako nejlepší. Kvalita algoritmů je ale často zásadně ovlivněna jejich parametry. Předpokládám, že pro vlastní algoritmus jste vykonal minimálně optimalizaci empiricky nastavených parametrů metodou pokusu a omylu, pokoušel jste se o podobnou optimalizaci i u ostatních algoritmů? Pokud ne, myslíte, že by některé z nich mohli zlepšit svoji kvalitu výsledků po optimalizaci svých nastavení?

2. Bylo by možné Vámi navrhnutou optimalizační metodu s úhlovou modulací také využít k optimalizaci analýzy sentimentu?

2. Could your proposed optimization method with angle modulation also be used to optimize sentiment analysis?

Conclusion

Ing. Raphael Kwaku Botchway's results meet the criteria set for a dissertation defense. The thesis has met the objective, is comprehensive, contains original elements and contributes new insights, and meets the criteria for a dissertation in Engineering Computer Science. The student has demonstrated creative skills in the given research area.

I recommend the thesis for defense, based on the submitted thesis I suggest Ing. Raphael Kwak Botchway to be awarded the academic degree of "philosophiae doctor" (abbreviation PhD.) after successful defense.

Trnava, 8 September 2023

Prof. RNDr. Jiří Pospíchal, DrSc.

Review of the Ph.D. thesis

„Soft Computing Techniques for Sentiment Analysis"

by Ing. Ralphael Kwaku Botchway

In his dissertation, the author focuses on the comments of social network users, which, with sophisticated processing, can then help companies to effectively target potential customers. The complexity of such a task, however, lies in the fact that the data is extremely large in number and, moreover, it does not have a uniform format and is not structured in any way. Another complication is that such outputs display emotions that amplify the meaning of the message and are very difficult to identify from the textual record or to eliminate in processing so that the results are not significantly distorted.

It is obvious that the processing of such large-scale data is burdened with a significant degree of uncertainty, there is no clear way to work with them, the focus of the thesis is so non-trivial, in terms of importance, topical and **meets the requirements for a doctoral dissertation**. At the same time, in addition to its theoretical contributions, it has the potential to be applied in the practice of companies with consequent economic benefits.

Natural language processing (NLP) in the context under study is called *sentiment analysis*, or *opinion mining*. To this, it is necessary to have a suitable metric for the semantic orientation of adjectives. VADER (Valence Aware Dictionary and sEntiment Reasoner) values represent the probability of a sentiment being *strongly positive, positive, neutral, negative*, and *strongly negative*.

Machine learning techniques, e.g., Naive Bayes, Maximum Entropy and K-Nearest Neighbour, are used for sentiment analysis tasks, but the author chooses a different approach based on metaheuristics and their combination.

In the introductory sections, he divides metaheuristic algorithms into four groups – *evolution-based,* swarm intelligence-based, *human behaviour-based,* and *physics-based algorithms* – and presents a comprehensive overview of *bio-inspired* algorithms. For the actual solution of the sentiment analysis, he chooses the *Particle Swarm Optimization* (PSO)

Given the No-Free-Lunch Theorem, one of the main objectives of this dissertation was to propose a solution based on a combination of metaheuristics, i.e. a memetic algorithm, to enhance the performance of the algorithms involved. The author satisfied this goal by designing a new Angle Modulated-based metaheuristic memetic algorithm combining Angle Modulated Particle Swarm Optimization (AMPSO) and Grey Wolf Optimization (GWO). The proposed algorithm was then validated on benchmarks from the University of California and the computational results confirmed its advantages compared to competing approaches. This satisfied another key objective of the dissertation. The fulfilment of the first of these goals is the main theoretical contribution of the dissertation, and the fulfilment of the second goal is then a confirmation of the quality of the proposed solution and at the same time a practical contribution.

The author also deals with the *feature selection problem* that has two competing goals: minimizing the number of features by removing redundant features and maximizing the classification accuracy. To realize both objectives, the author adopted the fitness function, expressed by Equation (27).

The Ph.D. thesis is logically structured into chapters and subchapters. The theoretical part was concisely explained and the practical part was presented with computational results, illustrative graphs and their detailed analysis.

The proposed algorithm was implemented in Python using open-source libraries and the partitioned data was repeated 30 times with statistical metrics, such as Average Classification Accuracy, Average Feature Selection Size, Mean Fitness Function, Best Fitness Function, Worst Fitness function, and Average Computational Time.

The language of the Ph.D. thesis and its typographical standard are very good, the text contains almost no bad formulations or misprints. I have only found the following ones:

Page 30: "These are computationally expensive methods than the filter" ... there should probably be "more expensive methods than".

Page 31: "Mirjarlili" – "Mirjalili"

Page 41: "Eighteen (18)" – "18" is unnecessary here.

Page 80: "so are its uses in the business world" – "its uses ... are ..."

Formal comments:

- Page 56: "computational complexity ... is $O(n\times(\mathcal{M}+\mathcal{M}\times\mathcal{B}+1)$" – expression can be simplified by neglecting the constant and slower growing terms on $O(n\times(\mathcal{M}\times\mathcal{B})$.

- It is rather strange that the author writes about himself in the 3rd person throughout the text, e.g.: "author of this thesis presents"

Questions:

1. Parsing texts on social media can be complicated by distorted words and grammatical errors. How do you work with "unrecognisable" parts of text? Are they disregarded or replaced with similar terms, as Google does, e.g., when inaccurate search queries are used?

2. On page 37 you say: "The primary goal of the PSOGWO (hybridization of PSO with GWO algorithm) is to increase the strength of both variants (i.e., PSO and GWO) by enhancing the GWO's exploration and PSO's exploitation capabilities." Explain in more detail the terms "exploration" and "exploitation" in this context.

3. Your work focuses on the social network Twitter (now X) and the banking services environment. Is this choice representative enough for sentiment analysis? (In our country, it is mainly used by a narrow layer of politicians)

4. Is it legal to download data from social networks? Or it is necessary to first obtain the consent of the operator of this network.

5. Will your approaches in social media data processing be used in the practice of specific companies?

Conclusion:

In my opinion, Ing. Ralphael Kwaku Botchway has proved to be capable of solving difficult research problems and a high level of programming abilities. He has designed a new memetic algorithm, the quality of which was confirmed by computational results.

The Ph.D. thesis satisfies conditions of the Czech Act 111/1998 and its Section 47, parts of the thesis have been published in 11 papers (3 in journals with impact factor), and therefore

I recommend

Ralphael Kwaku Botchway's Ph.D. thesis to be accepted by the Committee to be presented and defended in the Engineering Informatics study branch.

Brno, 11th September 2023

Prof. RNDr. Ing. Miloš Šeda, Ph.D.
Institute of Automation and Computer Science
Faculty of Mechanical Engineering
Brno University of Technology