An Improved K-Means Clustering with Machine Learning Based Sentiment Analysis and Classification Model

M. Ilayaraja
Department of Computer Science and Information Technology, Kalasalingam Academy of Research and Education, Krishnankoil, India.
Email: ilayaraja.m@klu.ac.in

Abstract— ***Sentiment analysis (SA) involves the task of automatically extracting the sentiments from user reviews via Natural Language Processing (NLP), data mining, and machine learning (ML) models. A major intention of SA is to identify the opinion and sentiments. It becomes helpful in the decision making of the customers whether to purchase an item or not. This paper develops an improved K-means clustering with random forest (RF) classification model called IKC-RF for effective SA. In order to handle the huge amount of online product reviews, K-means clustering technique is utilized to cluster the sentiments into appropriate class labels. Initially, the online product reviews are preprocessed and feature extraction process takes place. Next, clustering process is carried out by K-means clustering and finally, classification is done by RF technique. The application of clustering technique helps to handle the massive increase in dataset. The performance of the IKC-RF model is evaluated and the results are examined under distinct aspects.***

Index Terms— **Sentiment analysis, Sentiment Classification, K-means, Random forest, machine learning.**

I. INTRODUCTION

Sentiment analysis (SA) or opinion Mining (OM) is the model of automatic mining or classification of sentiments. They are exchangeable and they signify to similar denotation. However, a few research works illustrate that these 2 terms are somewhat several from everyone [1]. OM extract and examine the view on entities as SA searches the views, determines the sentiments illustrated and at last classifies the polarity. It is helpful in many application areas. In present days, the end user depicts their sentiment utilizing comments or analysis using internet. These evaluates are utilized and analyzed by main companies for decision making. Most of the online shopping sites (e.g. Amazon, Flipkart) support the users for giving feedback as sentiments on several characteristics of the purchased product.

A retailer also uses these reviews for enhancing the quality of the product and also to recognize the customer satisfaction. Other users create utilize of these reviews for making decisions whether to purchase the product or not [4-6].

The data sets utilized in SA roles a vital function. There are several methods for collecting data for SA. A main source of data is gathered from the product reviews. It can be useful to retailers as they can alter product quality depends on the SA outcomes. The most important review sources are the review websites. It is not only utilized to products, it is also helpful to stock exchanges [7], news [8] and debates of political parties [9]. From political debates, the people's view on election candidates is defined. Social media and blog are also a good source of data because of the nature of data distributing of people. Some of the benchmarked datasets are also accessible for performing study. In current days, further count of applications and improvements on SA methods are proposed. Massive count of researches is performed and count of articles is stays on enhancing day by day. In 2 elaborate surveys [9, 10] are projected that focuses only on the function and problems in SA. The techniques to overcome the problems are also discussed. Some short surveys are also projected and establish in the literature [11]-[13].

This paper develops an improved K-means clustering with random forest (RF) classification model called IKC-RF for effective SA. In order to handle the huge amount of online product reviews, K-means clustering technique is utilized to cluster the sentiments into appropriate class labels. Initially, the online product reviews are preprocessed and feature extraction process takes place. Next, clustering process is carried out by K-means clustering and finally, classification is done by RF technique. The application of clustering technique helps to handle the massive increase in dataset. The performance of the IKC-RF model is evaluated and the results are examined under distinct aspects.

II. IKC-RF Based SA for Online Product Reviews

The step contained in the SA of online product reviews utilizing improved RF (RF-K) method is listed under.

- Preprocessing of the online product reviews
- Extraction of features
- RF based SA with K-means (RF-K) model.

A. *Preprocessing*

The unwanted noise elements such as URLs, stop words, hash tags, multiple spaces and so on, requires to be eliminated in preceding to extracted features. The URLs are eliminated utilizing regular expression matching. The hash tag (#), punctuation marks such as /, _, \is removed and further count of white spaces is returned by a white space. Then, every word in the online reviews is changed to lowercase. The stop words (is, a, the, an and so on) and the words does not begin with the alphabets are also removed. Stop word dictionary [19] and acronym dictionary [20] are also utilized for improve the accuracy of the data set.

B. *Feature Extraction*

The pre-processed online product reviews are changed to feature vectors by calculating 10 features from the applied dataset. The extract features from the dataset are: Entire count of characters, Positive Emoji, Negative Emoji, Neutral Emoji, Positive Exclamation, Negative Exclamation, Negation, Positive Words, Negative Words, and Neutral Words.

C. *Clustering with Classification*

The normalized feature vector is performed to RF-K model that incorporates K-means and RF model to cluster the data. As K-means is the easiest and effective clustering model, it can be extremely utilized in many domains. However, it endures from the disadvantages of trapping to primary clusters. The generated cluster is utilized to further examination. Thus, in this case, the generated clusters from K-means are utilized in RF model for optimizing the cluster-heads. The count of online product reviews is enhancing day by day and the dataset size becomes enormous. As, the larger dataset can enhance the burden of RF and stuck to less classification accuracy. So, the presented model modifies the primary model of RF technique that leads to faster convergence and optimal classification accuracy. In RF-K model, the solutions achieved from K-means are performed to initialize of RF-K model.

Assume n refer the count of online product reviews that are clustered to N classes. All online products review is indicated by a feature vector that holds S count of features and all features has been scaled in [0, T]. The probability distribution of all features is calculated as follows:

$$p_i = \frac{O_i}{n} \qquad (1)$$

where I refer ith feature value $(0 \leq i \leq T)$ and O_i denotes the entire count of online product reviews has ith feature value.

$$\mu = \sum_{i=1}^{T} i\,p_i \qquad (2)$$

Some online product review is classified to class D_j for that it has minimal Euclidean distance. Thus, the probability of incidence w_j of class is equated as,

$$w_j = \sum_{i \in D_j} p_i \qquad (3)$$

The mean of class $D_j$is calculated as

$$\mu_j = \sum_{i \in D_j} \frac{i p_i}{w_j} \qquad (4)$$

The inter-class variance is usually determined as:

$$\sigma^2 = \sum_{j=1}^{N} w_j (\mu_j - \mu)^2 \qquad (5)$$

For the clustering model of several online product reviews, the inter-class variance depicted in Eq. (5) must be maximized. RF classifier is a novel and important tree-based method that based on the integrated on the tree of predictors, in order that all trees is dependent on values of arbitrary vector that suffers sampling independently and with the same distribution to all trees called as RF. It contained integration of separate base classifiers, in which every tree is established utilizing an arbitrary vector sampled in an independent method from the classifier input vector to activate a rapid production of tree. For classifying data, the classification individual vote from every tree is integrated with the support of the functional rule based model.

An effective RF is created utilizing the decision trees (DT) attained from the forest. It functions on 2 levels. An initial level creates a tree with the help of the samples selected in an arbitrary method. All trees in the forest endure training by utilize of distinct instances of similar sizes. Only partial training samples are utilized for training the trees and remaining are executed to cross validation method. It can be complete for determining the outcomes of the RF classifier. It can be depending on the model of "robustness" of trees is balanced even at the time of declining connection among the trees. Then, divide constraints to all nodes in the tree are shared as predictor variables. An important process is to select the count of variables that gives minimum connection with adequate predictive strength. Thus, an optimum range is achieved to the subset of predictive attributes that is usually utilized to optimum test subset. Currently, the group of 2 existing parameters is contained in the RF classifier. It creates utilize of GINI index to the computation of the impurities of parameters interms of classes. The GINI index is defined utilizing the following function:

$$\sum\sum_{j \neq i} \left(\frac{f(C_i,T)}{|T|}\right)\left(\frac{f(C_j,T)}{|T|}\right) \qquad (6)$$

where T refers the provided training dataset, C refers the class of arbitrarily chosen case that falls into $f(C_i,T)$, contains the possibility of selected values that goes to C_i. When the defined GINI index value is improved, afterward the heterogeneity of classes obtains enhanced. But, in case of the minimization in GINI index, afterward the enhanced, after that class heterogeneity improves. Although, as there is a decrease in GINI index, next heterogeneity of classes obtains increased. If a child node of GINI index is smaller over parent node, after that following split achieved is efficient. Hereafter, tree dividing is ended at the time of GINI index attaining zero value that signifies the single class exists in all nodes in tree. When all trees are developed in the forest utilizing above reasons, afterward classifying model occurs by utilize of an efficient dataset.

As said, to insert images in *Word*, position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with “Float over text” unchecked).

The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

III. Math

If you are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

IV. Units

Use either SI (MKS) or CGS as primary units. (SI units are strongly encouraged.) English units may be used as secondary units (in parentheses). **This applies to papers in data storage.** For example, write "15 Gb/cm^2 (100 Gb/in^2)." An exception is when English units are used as identifiers in trade, such as "3½ in disk drive." Avoid combining SI and CGS units, such as current in amperes and magnetic field in oversteps. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity in an equation.

The SI unit for magnetic field strength *H* is A/m. However, if you wish to use units of T, either refer to magnetic flux density *B* or magnetic field strength symbolized as $\mu_0 H$. Use the center dot to separate compound units, e.g., "A·m^2."

V. Helpful Hints

A. Figures and Tables

Because the final formatting of your paper is limited in scale, you need to position figures and tables at the top and bottom of each column. Large figures and tables may span both columns. Place figure captions below the figures; place table titles above the tables. If your figure has two parts, include the labels "(a)" and "(b)" as part of the artwork. Please verify that the figures and tables you mention in the text actually exist. **Do not put borders around the outside of your figures.** Use the abbreviation "Fig." even at the beginning of a sentence. Do not abbreviate "Table." Tables are numbered with Roman numerals.

Include a note with your final paper indicating that you request color printing. **Do not use color unless it is necessary for the proper interpretation of your figures.** There is an additional charge for color printing.

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity "Magnetization," or "Magnetization *M*," not just "*M*." Put units in parentheses. Do not label axes only with units. As in Fig. 1, for example, write "Magnetization (A/m)" or "Magnetization (A $\cdot$ m^{-1})," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

Multipliers can be especially confusing. Write "Magnetization (kA/m)" or "Magnetization (10^3 A/m)." Do not write "Magnetization (A/m) × 1000" because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 8 to 12 point type.

B. References

Number citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2]. Multiple references [2], [3] are each numbered with separate brackets [1]–[3]. When citing a section in a book, please give the relevant page numbers [2]. In sentences, refer simply to the reference number, as in [3]. Do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] shows ... ." Number footnotes separately in superscripts (Insert | Footnote).[1] Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

Please note that the references at the end of this document are in the preferred referencing style. Give all authors' names; do not use "*et al*." unless there are six authors or more. Use a space after authors' initials. Papers that have not been published should be cited as "unpublished" [4]. Papers that have been submitted for publication should be cited as "submitted for publication" [5]. Papers that have been accepted for publication, but not yet specified for an issue should be cited as "to be published" [6]. Please give affiliations and addresses for private communications [7].

C. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations such as SI, ac, and dc do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.," not "C. N. R. S." Do not use abbreviations in the title unless they are unavoidable (for example, "INTERNATIONAL JOURNAL OF ENGINEERING AND INNOVATIVE TECHNOLOGY" in the title of this article).

D. Equations

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First use the equation editor to create the equation. Then select the "Equation" markup style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$\begin{aligned}\int_0^{r_2} F(r,\varphi)\,dr\,d\varphi &= [\sigma r_2/(2\mu_0)] \\ &\cdot \int_0^{\infty} \exp(-\lambda\,|z_j - z_i|)\,\lambda^{-1}\,J_1(\lambda r_2)\,J_0(\lambda r_i)\,d\lambda\,.\end{aligned} \tag{1}$$

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit tesla). Refer to "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ... ."

E. Other Recommendations

Use one space after periods and colons. Hyphenate complex modifiers: "zero-field-cooled magnetization." Avoid dangling participles, such as, "Using (1), the potential was calculated." [It is not clear who or what used (1).] Write instead, "The potential was calculated by using (1)," or

Accepted 31st August 2020

"Using (1), we calculated the potential."

Use a zero before decimal points: "0.25," not ".25." Use "cm^3," not "cc." Indicate sample dimensions as "0.1 cm × 0.2 cm," not "0.1 × 0.2 cm^2." The abbreviation for "seconds" is "s," not "sec." Do not mix complete spellings and abbreviations of units: use "Wb/m^2" or "webers per square meter," not "webers/m^2." When expressing a range of values, write "7 to 9" or "7-9," not "7~9."

A parenthetical statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.) In American English, periods and commas are within quotation marks, like "this period." Other punctuation is "outside"! Avoid contractions; for example, write "do not" instead of "don't." The serial comma is preferred: "A, B, and C" instead of "A, B and C."

If you wish, you may write in the first person singular or plural and use the active voice ("I observed that ..." or "We observed that ..." instead of "It was observed that ..."). Remember to check spelling. If your native language is not English, please get a native English-speaking colleague to proofread your paper.

VI. Some Common Mistakes

The word "data" is plural, not singular. The subscript for the permeability of vacuum μ_0 is zero, not a lowercase letter "o." The term for residual magnetization is "remanence"; the adjective is "remanent"; do not write "remnance" or "remnant." Use the word "micrometer" instead of "micron." A graph within a graph is an "inset," not an "insert." The word "alternatively" is preferred to the word "alternately" (unless you really mean something that alternates). Use the word "whereas" instead of "while" (unless you are referring to simultaneous events). Do not use the word "essentially" to mean "approximately" or "effectively." Do not use the word "issue" as a euphemism for "problem." When compositions are not specified, separate chemical symbols by en-dashes; for example, "NiMn" indicates the intermetallic compound $Ni_{0.5}Mn_{0.5}$ whereas "Ni–Mn" indicates an alloy of some composition Ni_xMn_{1-x}.

Be aware of the different meanings of the homophones "affect" (usually a verb) and "effect" (usually a noun), "complement" and "compliment," "discreet" and "discrete," "principal" (e.g., "principal investigator") and "principle" (e.g., "principle of measurement"). Do not confuse "imply" and "infer."

Prefixes such as "non," "sub," "micro," "multi," and ""ultra" are not independent words; they should be joined to the words they modify, usually without a hyphen. There is no period after the "et" in the Latin abbreviation "*et al.*" (it is also italicized). The abbreviation "i.e.," means "that is," and the abbreviation "e.g.," means "for example" (these abbreviations are not italicized).

An excellent style manual and source of information for science writers is [9].

VII. Editorial Policy

The submitting author is responsible for obtaining agreement of all coauthors and any consent required from sponsors before submitting a paper. It is the obligation of the authors to cite relevant prior work.

Authors of rejected papers may revise and resubmit them to the journal again.

VIII. Publication Principles

The contents of the journal are peer-reviewed and archival. The journal INTERNATIONAL JOURNAL OF ENGINEERING AND INNOVATIVE TECHNOLOGY (IJEIT) publishes scholarly articles of archival value as well as tutorial expositions and critical reviews of classical subjects and topics of current interest.

Authors should consider the following points:

1) Technical papers submitted for publication must advance the state of knowledge and must cite relevant prior work.
2) The length of a submitted paper should be commensurate with the importance, or appropriate to the complexity, of the work. For example, an obvious extension of previously published work might not be appropriate for publication or might be adequately treated in just a few pages.
3) Authors must convince both peer reviewers and the editors of the scientific and technical merit of a paper; the standards of proof are higher when extraordinary or unexpected results are reported.
4) Because replication is required for scientific progress, papers submitted for publication must provide sufficient information to allow readers to perform similar experiments or calculations and use the reported results. Although not everything need be disclosed, a paper must contain new, useable, and fully described information. For example, a specimen's chemical composition need not be reported if the main purpose of a paper is to introduce a new measurement technique. Authors should expect to be challenged by reviewers if the results are not supported by adequate data and critical details.

IX. Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank ... ." Instead, write "F. A. Author thanks ... ." **Sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page**.

REFERENCES

[1] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[2] J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[3] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, “Rotation, scale, and translation resilient public watermarking for images,” IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

Accepted 31st August 2020