TEXT CLASSIFICATION

Yasaman Mirmohammad*

Computer and IT Department,Amirkabir University of Technology,Tehran,Iran

E-mail: ys.m@aut.ac.ir

Abstract

The classification has been studied widely in data mining and machine learning with applications in different areas such as Recommender systems, target marketing, disease trajectory, law, Spam Filtering, Sentiment Analysis, etc. In this document, I will provide a survey of text classification techniques employed in practice, their applications, strengths and weaknesses, and current research trends to provide improved awareness regarding knowledge extraction possibilities.

Introduction

There has been an exponential increase in the number of complex documents that require a deeper understanding of machine learning methods to be able to accurately classify texts in many applications[1] Structured data is the most significant need for every field, such as universities, businesses, research institutions, government projects, and technology companies.[2] Since eighty percent of data about an entity can only be found in the Unstructured form,[2] this task has been one of the most challenging topics of today's researches. Most of the data we can provide has the information implicitly in itself, so Text mining tries to discover the hidden non-linear relationships in a dataset to gain useful patterns that illustrate

the knowledge contained in the data.[3] Text analytic approaches try to model these relationships and patterns by converting text to numbers.[4] Text classification system contains a pipeline:

- **Feature Extraction**[5] [6]
- **Dimensionality Reduction**

  The most frequently used methods in text analysis such as Principle component analysis(PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), Non-Negative Matrix Factorization (NMF), Random Projection, Random Kitchen Sinks, Johnson Lindenstrauss Lemma, and Autoencoder have been explained in.[1]
- **Classification Techniques**[7] [8] [9]
- **Evaluation**

  Accuracy calculation is the simplest method of evaluation but does not work for unbalanced data sets.In this situation,other substitutes like FB score,AUC(are under curve) and ROC(receiver operating characteristics) could be used.[1]

  While there are various methods introduced, Clustering, Classification, and Categorization are three primary techniques followed in text analytics.[10]

Classification

The problem of classification is defined as follows. We have a set of training records D={X1, . . . , XN} such that each record is labeled with a class value drawn from a set of k different discrete values indexed by {1 . . . k}. The training data is used in order to build a classification model, which relates the features in the underlying record to one of the class labels is used to predict a class label for this instance. In the hard version of the classification problem, a particular label is explicitly assigned to the instance. In contrast, in the soft version of the classification problem, a probability value is assigned to the test

instance. Other variations of the classification problem allow ranking of different class choices for a test instance or allow the assignment of multiple labels to a test instance.[11]

The problem of text classification is closely related to that of classification of records with set-valued features.[12] Nevertheless, the frequency of words also plays an instrumental role in the text classification process, and the typical size of text data is much higher than a typical set-valued classification problem.[11] A complete survey of a variety of classification methods can be found in[13],[4] and a survey that is focused on the text domain can be found in.[14] An evaluation of different kinds of text classification methods can be found in.[15]

Statistical topic modeling is applied for multi-label document classification, where each document gets assigned to one or more classes. It became a hot topic in the past decade since it performed very good for datasets with massive instances for an entity.[16] Unsupervised learning helps in the field of big data.[17]

As mentioned above, computational complexity also increases when documents become larger.[18] Machine learning employs advanced models to make decisions based on its cognizance, unlike statistics.[19] [20]Since a purely statistical and purely machine learning approach is considered to be less competent, a hybrid approach is preferred (Srivastava, 2015). Logistic regression is an efficient probability-based linear classifier for overcoming the overfitting problem(means that the model memorizes the dataset's patterns instead of the learning procedure, in fact, an over-generalization occurs) by using penalized logistic regression in active learning algorithm.[21]

Multi-valued and multi-labeled data makes it challenging to choose a particular set of attributes or to calculate similarity scores of each.[22] The decision tree algorithms calculate the similarity scores comprehensively and accurately. To overcome the problem from the order of classes in rule learning, Complexity-based Parallel Rule Learning algorithm is suggested.[23]

Instance selection technique helps knowledge discovery procedure. An instance selector based on Support Vector Machine (SVM)is suggested to reduce the amount of data by

filtering out noise from a given training dataset.[24] The multi-class classification has been combined kernel density estimation with kNN[25] that improves the weighting principle of k-NN and increasing the accuracy of classification. It has been seen as efficient for complicated classification. Artificial Neural Networks can solve the problem of high dimensional and large data. Fuzzy adaptive approaches are scalable.[26]

This literature review aims to analyze various text classification techniques with their strengths and weaknesses and to provide improved awareness regarding various knowledge extraction possibilities in the field of data mining and text analysis.

Research Overview

The methods have been illustrated in a tree structure after analyzing the similarities and differences among the various approaches. (figure1). This branching has been done due to it is simplicity and generalization, concerning the structure indicated in.[4] The various search terms used in Google scholar were: text classification, text + classification, text + classification + review, and all the subheadings stated in Figure 1. I have searched for articles between 2000,2019 mostly.

Text Classification

Usages and importance

Information retrieval systems, Information Filtering, Sentiment Analysis, Recommender Systems, Knowledge Management, Document Summarization, Marketing,etc. are the most widespread areas of application.[1]

Generally, a classification technique could be divided into two approaches: statistical and machine learning. Statistical techniques purely satisfy the proclaimed hypotheses manually. Therefore the need for algorithms is little. However, Machine Learning techniques

were specially innovated for automation.[27] Figure 1 shows this difference better. Supervised Learning(Classification for discrete features and Regression for continuous features) covers the tasks in which there exists label in the gathered data. Among the supervised classification algorithms, there are two types: parametric and non-parametric, based on the supremacy of parameters in the data.K-NN, Support Vector Machine (SVM), Decision Tree, Rule Induction and Neural Networks are the most significant non-parametric methods.[28] Logistic Regression and Naive Bayes are the most utilized parametric classification algorithms.[29] Graph-based methods,co-training, self-training, and transductive SVM are some of widely used semi-supervised learning methods.[30] [31]

Clustering(fuzzy c-means,k-means, and Hierarchical)is the unsupervised learning approach since there is no label in data while creating the model. Below are some of the text classification techniques and their research directions

In,[27] Classification algorithms have been analyzed from the aspect of text mining.

Machine Learning Approaches

The significant increase in data size, velocity, has made the automation essential in text processing methods, including text classification.

In some situations, defining a set of logical rules using knowledge-engineering techniques and based on expert opinions to classify documents helps to automate the classification task. Text classification could be divided into three categories: supervised, semi-supervised and unsupervised, based on the learning strategy followed by the data model in.[32]

In machine learning terminology, the classification problem comes under the supervised learning principle, where the system is trained and tested on the knowledge about classes before the actual classification process. Unsupervised learning occurs when labeled data is not available or accessible. Since this process is complicated and has performance issues, It is more suitable for big data. Semi-supervised learning is followed when data is only partly labeled.[33] Establishing an accurate relationship between labeled and unlabeled data is not

so simple. The efficiency is measured using metrics like Accuracy, Precision, and Recall. When the dataset is enormous, the classification errors tend to be less. It has also been known that the selection of suitable algorithms for a particular dataset plays a significant role in text classification's quality.

Unsupervised learning

Unsupervised learning is a type of Machine Learning algorithm in which instances are drawn from the data by clustering data into different categories without labeled responses (expected outcomes). No training data is provided. As more data is fed into the model, the algorithm refines itself to efficiency. Clustering and Principal component analysis(PCA) are frequently used in unsupervised learning. In many scenarios, clustering is the same as unsupervised learning. In many cases, expert knowledge required to label the instances is either non-existent or insufficient. In such a situation, self-organizing maps and correlation coefficient are used to cluster the documents and to label the documents for further classification.[34] This method eliminates the curse of dimensionality (various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience) and expert intervention as well. This kind of hybrid model is more suitable for massive data. Feature extraction in high dimensional data can be done via statistical cluster analysis. It also reduces the time and cost complexity of the pre-processing procedure.[35] Query type classification detects the categories of search queries and labels them.[36] Transactional query classification is a problem of unsupervised learning approach.[37] A Transaction(what a user executes after a search engine returns the queried data) is a query that can be converted to patterns to derive knowledge about the information behind the queries. In this way, the web forms can be used to optimize search engine performance by ranking the search results shortly. In document clustering, the number of data points, dimensionality, and the number of clusters would increase with time, so the algorithms should have enough space for expan-

sion in such cases. Filtering untrustworthy data from data sources is an exciting option for future researches in such cases.[38] Some of the famous unsupervised learning algorithms are principal component analysis(PCA), independent component analysis, anomaly detection, Hebbian learning, expectation-maximization algorithm, singular value decomposition(SVD), non-negative matrix factorization, and all those mentioned in Figure 1.[39]

Semi-Supervised learning(SSL)

Semi-supervised learning is a combination of supervised and unsupervised learning techniques. This type of learning employs a small amount of labeled data and a large amount of unlabeled data for training. The labels are assigned by combining labeled and unlabeled instances, as unlabeled data mitigate the effect of insufficient labeled data on classifier accuracy. Some of the SSL techniques include, self-training or bootstrapping, co-training, transductive SVMs, generative models and graph-based methods and Vector space models are mostly used in language processing problems to address natural language semantics that supposes words in similar contexts have similar meanings. Meaning values are calculated according to the Helmholtz principle. This model is non-iterative but effective in augmenting the efficiency of the classifier. The system can be combined with semantic kernels that smooth document term vectors using term to term semantic relations. Finding out more approaches to extract the information from the context of a class could be tried in the future.[4]

Traditional text classification approaches do not work when there is no labeled data for a particular class of the dataset; for example, if the labeled data is only available only for positive samples, A semi-supervised algorithm based on tolerance roughest and ensemble learning is recommended for this issue.[40] The unavailable class is extracted approximately from the dataset and set as the labeled sample. The ensemble classifier iteratively builds the margin between positive and negative classes to further approximate harmful data, since negative data is mixed with the actual data. Therefore, without the need for training samples, classification is achieved through a hybrid approach. It eliminates the cost of hand labeling

data, especially in big data. The application of semi-supervised algorithms is highly useful in information filtering requirements.[41] The role of semi-supervised algorithms in multi-label hierarchical classification is an area where there is still a need for more exploration. Self-training, along with a semi-supervised classifier, is recommended for multi-label hierarchical classification. It has also proven a better way to achieve automatic label attribution.

Supervised learning

Supervised learning is the most expensive and highly difficult type of Machine Learning approaches. The main reason behind this notion is that it requires a human intervention while assigning labels to classes, which is not possible in large datasets. Though the workflow mimics the techniques followed in AI processes, it is time-consuming. It is also called inductive learning in ML.[37] Supervised learning becomes expensive when different data distributions, different outputs, and different feature spaces occur as in heterogeneous text corpora. One of the most widely used supervised methods is maximum likelihood estimation(MLE).[42] Prior assumptions could simplify the learning process. These kinds of assumptions about data introduce two approaches, such as parametric and non-parametric.

- **Parametric Approaches**

  A parametric model can summarize data based on underlying parameters.[43] The most widely used parametric classifiers are Logistic Regression, Naive Bayes, and also Rocchio Classification.[44] Bagging, Boosting could also be parametric their base classifier is parametric.

  - **Rocchio Classification**

    Using relevance feedback to query full-text databases is the base of this algorithm. It has been applied in different applications since 1971 for document categorization.[45]

This classification algorithm uses TF-IDF weights for words instead of boolean features.[1]

– **Logistic Regression**

One of the most basic parametric classification algorithms is logistic Regression (LR) which has been addressed in most data mining domains.[1] Logistic Regression focuses on selecting the best subjects to be labeled to achieve a functional classification. It is an opportunity to reduce temporal costs. Active learning is employed to find the best subjects to the label in ML models, which is a growing field of research in text mining. It has proven to minimize the generalization error of models. Auto adapting regularization parameters and applying a penalized logic regression-based active learning to multi-class problems is suggested for future research.[46] Kernel methods transform data into higher dimensional space in contrast to linear classifiers that are implemented directly on data in its original space. The imbalanced data has been a problem. A new logistic regression based on rare event weights Kernel is recommended. It is also easier to implement even if there are a massive data imbalance and numerous rare events.[47] Linear classifiers are suitable for large and high-dimensional datasets.[4] In the future, these kinds of tasks could be extended to evaluate the relationship between sentences rather than words.[48] The automatic text categorization is the process of assigning textual documents to predefined categories based on defined or undefined contents. However, we face an issue when the number of features exceeds the number of observations. Also, ML techniques tend to perform weakly due to these overfitting problems.[48] Usually for coming over this issue, different methods are utilized: Ensembling (combining predictions from multiple separate models), cross-validation (Use your initial training data to generate multiple mini train-test splits and then Use these splits to tune your model), Early stopping (stopping the training process before the learner passes that point of overfitting), Removing fea-

tures, training with more data, and Regularization (discourages learning a more complex or flexible model by adding hyper-parameter to the function, so as to avoid the risk of overfitting).

One innovated approach for solving the overfitting problem is controlling the complexity of the model, during the training process, using model selection techniques. It is shown that Logistic Regression is better suited for these kinds of problems than SVMs.[49] [4] Two well-known Regularization methods are used: Ridge and lasso. Ridge logistic regression is a popular solution to the text categorization problem; however, its role in large scale documents is still questionable.[50] To eliminate this difficulty, the sparse solution is combined with ridge regression. The sparsification removes less important features there-by solving the severe problem of ridge regressors.[50]

– **Naive Bayes Classifier**

Probabilistic classifiers are commonly used in learning extensive data by using a Multinomial model.

The performance could be enhanced by linking the dependencies among attributes. Attributes play an essential role in classification, and the algorithm improves by linking the dependencies among the features. This method is often used in preprocessing data for the simplicity of computation.[4] Assigning different weights to attributes can highly improve the performance of the model.[51] Deep feature weighting is also useful.[52] [4]

The performance of Naive Bayes depends on the accuracy of the estimated conditional probability terms. In some cases, the advantages of Naive Bayes are challenged by a strong conditional independence assumption between attributes.[53] To sum up, Naive Bayesian classifiers are simple and powerful in It is computationally inexpensive and also needs a meager amount of memory.[54] [4]

- **Non-Parametric Approaches**

  The model that can not summarize data based on underlying parameters is called a non- parametric model. Neural networks, Support vector machines(SVM), k-nearest neighbor(K-NN), Decision trees, and rule induction are famous non-parametric classifiers that are opened up below.

  - **Neural Network models**

    Artificial neural networks (ANNs) work in the same way as the human brain in decision making.[55]Evolution algorithms are used to generalize a neural network model, with minimum or no human intervention.[4] Different types of Neural Networks widely used for the text data such as:Recurrent Neural Network(RNN),Deep neural networks(DNN),Long Short-Term Memory (LSTM),Gated Recurrent Unit (GRU),Convolutional Neural Networks (CNN),Deep Belief Network (DBN),etc are stated in.[1]

    Neural networks are popular among cases where a hierarchical multi-label classification approach is required.[4] This kind of classification is sophisticated as it is similar to the fuzzy clustering approach: each sample may belong to more than one class(as in fuzzy clustering, every sample may belong to more than one cluster), in this case, predictions of one level is fed as inputs to next level to make a final decision.[56] Also, we can avoid overfitting by Adaboost method, and improve the accuracy.[57]

    A comparison between traditional and deep learning techniques has been discussed in.[1]

  - **Support vector machines**

    The Support Vector Machine (SVM) algorithm is employed for various classification problems.[58] SVM was originally designed for binary classification tasks.[59] [60]It has its applications in credit risk analysis, medical diagnosis, text categorization,

and information extraction.[4] SVMs are more suitable for high dimensional data. The most significant feature in these algorithms is that the complexity of the classifiers will depend on the number of support vectors, instead of dimensions (number of attributes). So the same hyperplane is produced for all the training sets, and they are better in generalization.[61] SVM performs with the same accuracy when the data is sparse.[4]

Kernels increase the performance of the SVMs. It includes the background details for text categorization. Also, an effective learning method for text categorization based on SVM is suggested that reduces labeling effort by selecting appropriate samples for labeling, not all.[62]

SVM can be performed by semi-supervised clustering for text classification for determining the category of text from multiple components.[4] The unlabeled data is also used to improve performance. Though it is shown to outperform the traditional SVMs, it still lags in handling unbalanced data. An efficient instance selection based SVM is suggested for better result.[63] One class support vector machine is an accurate anomaly detection method for improving the accuracy of text classification problems.[64]

– **K-nearest neighbor**

(Li, L.; Weinberg, C.R.; Darden, T.A.; Pedersen, L.G. Gene selection for sample classification based on gene expression data: Study of sensitivity to the choice of parameters of the GA/KNN method. Bioinformatics 2001, 17, 1131–1142) K-Nearest Neighbor (k-NN)uses the closest training samples for determining classes(points that are closer to each other belong to one class). It is also called instance-based learning.[65] This is a kind of lazy algorithm(focusing on a generalization of the training data, in contrast with an eager algorithm). Since K-NN is a non-parameter model, it is too noise-sensitive, and the computational complexity increases with high dimensional data. Therefore, Deciding the value of k is

complicated. To reduce the cost of computing k, Tree-based k-NN is used, which reduces search scope, through better traversing techniques.[66]

k-NNs are also most famous for classifying instances based on the context of data through majority voting.[67] This method is especially suitable for small datasets.

The weight-adjusted k-nearest neighbor classification (WAKNN) is a version of KNN which tries to learn the weight vectors for classification.[68] [1]

KNN is easy to implement and adapts to any feature space. This model also naturally handles multi-class cases.[69]

– **Decision trees**

Decision trees are highly comprehensible models. C4.5 and CART, and ID3 are the most used decision tree techniques[70] Fuzzy ID3 is another popular variant that incorporates the fuzziness of attributes into decision rules. Ensemble-based trees(like random forest) use boosting and bagging techniques to combine more than one classifier that employs different decision rules for different datasets.[71] These ensembles have a remarkable performance compared to regular decision trees. However, computational cost increases as each input query are fed to every component classifier.[?][72] Streaming data is an essential challenge in the data processing arena.[4]

High dimensional data has been another challenge for decision trees. To solve this problem, cluster trees are suggested.[73]

The decision tree is fast for learning and prediction, but it is also susceptible to small perturbations in the data.[67]

– **Rule induction**

Classification of free text with minimal label description is a significant problem in text categorization. A rule-based framework of lexical, syntactic patterns is chosen as classification features that reduce common classification errors. In this approach, the performance is measured using a metric called sensitivity analysis,

which optimizes the number of rules that support efficient categorization. The rules are dependent on the lexicon input, which describes the domain of documents under consideration. Therefore, the categorization is more effective.[74] RIPPER is a famous rule induction technique. Rule order is optimized using the ant colony algorithm on the decision list. The decision list is mostly in the form of 'if, then and else if' structure. Genetic algorithm and Simulated annealing are other rule order optimization techniques widely used. Some of the significant disadvantages in this technique are the nature of rules being dependent on previously generated rules, and rule learning occurs sequentially. Selecting the best routing scheme for ordering the rules is another useful research direction.[75] [23]

Statistical Approach

Statistical Approaches are purely mathematical processes, and they act as the mathematical foundation for all other text classifiers. It works similar to a computer program, executing the given instructions without any ability of its own.[4] To achieve a good classification, the amount of information to be handled by the application has to be accurate, and it is achieved by reducing the dimensionality (number of features) in the data.[76] Data in emails are complex and multi-dimensional. Statistical feature extraction techniques, such as Principal Component Analysis (PCA), Biased Discriminant Analysis (BDA), and Average Neighborhood Margin Maximization (ANMM), have been proven to be better dimensionality reduction techniques.[77] They are ordered by relevance but not suitable for non-linear data. Statistical techniques are shown to be inefficient for large datasets. In the future, these methods can also be used for binary text classification, or spam filtering (to determine whether a particular email is a spam or not). Heavily correlated features play a significant role in binary text classification. In another word, similar to anomaly detection processes, on the lines of non-parametric classification trees, classifier filters unknown authors' text documents. It is based on the combination of Kruskal–Wallis and Brunner–Dette–Munk

test.[78] This model recognizes the most common words used by a known author and identifies new authors. It can be applied for fraud detection applications, too.

Conclusion

Based on the study carried out for this article, it has been found that the most widely used text classification techniques follow a semi-supervised learning approach.[4] Since, it has the potential to improve classification efficiency, by the combined benefits of both supervised and unsupervised learning techniques. It is also found suitable for solving the labeling problem while handling more number of instances of an entity. It is found that the active learning method is followed to reduce the temporal costs involved by selecting only the most suitable instance to classify a sample (iterative supervised learning)[79] The genetic algorithm helps to achieve optimal ordering of rules in the decision list. It is found to eliminate conflicts among the generated rules and improve the accuracy of the model.

Data warehouses play an essential role in any analysis. Data ingestion is the crucial phase in maintaining large datasets and accessing them for knowledge discovery. It takes two forms, such as batch processing and streaming ingestion.[80] It could also be scaled up using cloud technologies with little effort. Google BigQuery and Amazon Redshift are well-known and popular data warehouses with cloud support. Deep meaning-extraction, algorithm efficiency, semantic analysis, audit automation, data scalability, data breach, and real-time decision making are some of the areas in need for further research.[4] It is seen that the classification time taken by k-NN is increasing, and it is challenging to estimate optimal k value.[66] Though decision trees reduce complexity, one mistake will make the entire sub-tree go wrong.[81] Parameter tuning is a significant issue in SVMs ınel-surveyınel-surveyınel-survey.[?] Voting algorithms like boosting techniques are known for high accuracy; however, they require complicated calculations and high memory.[82] These indicate that there is no specific algorithm for a particular text classification problem concerning automation.[4]

The algorithms discussed in this study are summarized according to their strengths and weaknesses in[4] and more detailed in.[1]

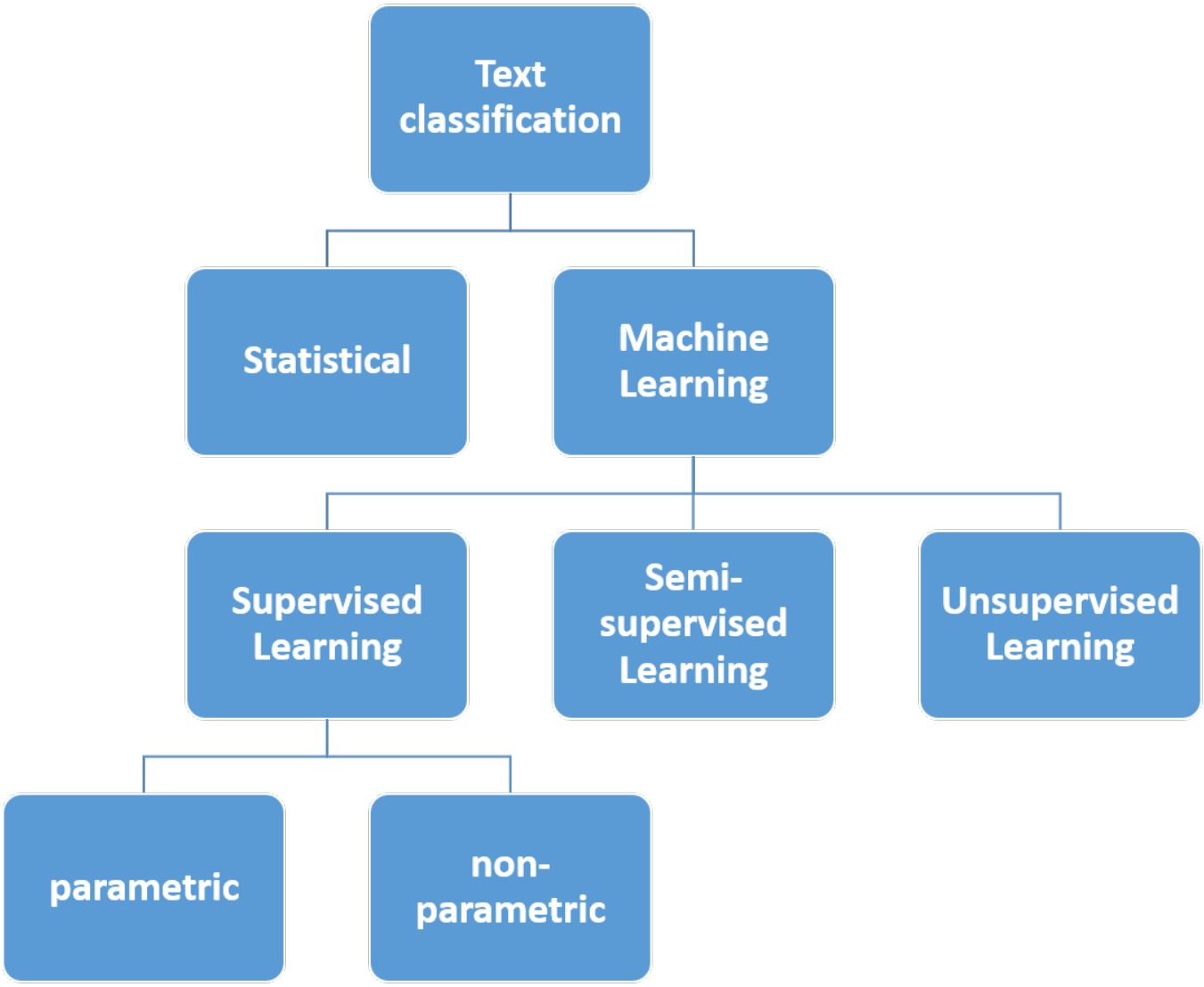

Figure 1: Representation of Text Classification methods

References

(1) Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. Information 10, no. 4 (2019): 150.

(2) Khan A, Baharudin B, Lee LH, Khan K. A review of machine learning algorithms for text-documents classification. Journal of advances in information technology. 2010 Feb;1(1):4-20.

(3) Brindha, S., Sukumaran, S., Prabha, K. (2016). A survey on classification techniques for text mining. Proceed- ings of the 3rd International Conference on Advanced Computing and Communication Systems. IEEE. Coimbatore, In-dia. https://doi.org/10.1109/ICACCS.2016.7586371.

(4) Thangaraj, M., and M. Sivakami. TEXT CLASSIFICATION TECHNIQUES: A LITERATURE REVIEW. Interdisciplinary Journal of Information, Knowledge Management 13 (2018).

(5) Goldberg, Y.; Levy, O.Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv 2014, arXiv:1402.3722.

(6) Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors forWord Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Volume 14, pp. 1532–1543.

(7) Mamitsuka, N.A.H. Query learning strategies using boosting and bagging. In Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98); Morgan Kaufmann Pub.: Burlington, MA, USA, 1998; Volume 1.

(8) Kim, Y.H.; Hahn, S.Y.; Zhang, B.T. Text filtering by boosting naive Bayes classifiers. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24–28 July 2000; pp. 168–175.

(9) Schapire, R.E.; Singer, Y. BoosTexter: A boosting-based system for text categorization. Mach. Learn. 2000, 39, 135–168.

(10) Vasa K. Text classification through statistical and machine learning methods: A survey. International Journal of Engineering Development and Research. 2016;4:655-8.

(11) Aggarwal CC, Zhai C. A survey of text classification algorithms. InMining text data 2012 (pp. 163-222). Springer, Boston, MA.

(12) Cohen WW. Learning trees and rules with set-valued features. InAAAI/IAAI, Vol. 1 1996 Aug 4 (pp. 709-716).

(13) Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley Sons; 2012 Nov 9.

(14) Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002 Mar 1;34(1):1-47.

(15) Yang Y, Liu X. A re-examination of text categorization methods. InSigir 1999 Aug 15 (Vol. 99, No. 8, p. 99).

(16) Rubin TN, Chambers A, Smyth P, Steyvers M. Statistical topic models for multi-label document classification. Machine learning. 2012 Jul 1;88(1-2):157-208.

(17) Zhoua, L., Pana, S., Wanga, J., Athanasios, V., Vasilakos. (2017). Machine learning on big data: opportunities and challenge. Neurocomputing, 237, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026.

(18) Staš J, Juhár J, Hládek D. Classification of heterogeneous text data for robust domain-specific language modeling. EURASIP Journal on Audio, Speech, and Music Processing. 2014 Dec;2014(1):14.

(19) Du JH. Automatic text classification algorithm based on Gauss improved convolutional neural network. Journal of computational science. 2017 Jul 1;21:195-200.

(20) Ranjan NM, Prasad RS. Automatic text classification using BPLion-neural network and semantic word processing. The Imaging Science Journal. 2018 Feb 17;66(2):69-83.

(21) Wang J, Park E. Active learning for penalized logistic regression via sequential experimental design. Neurocomputing. 2017 Jan 26;222:183-90.

(22) Yi W, Lu M, Liu Z. Multi-valued attribute and multi-labeled data decision tree algorithm. International Journal of Machine Learning and Cybernetics. 2011 Jun 1;2(2):67-74.

(23) Asadi S, Shahrabi J. ACORI: a novel ACO algorithm for rule induction. Knowledge-Based Systems. 2016 Apr 1;97:175-87.

(24) Tsai CF, Chang CW. SVOIS: support vector oriented instance selection for text classification. Information Systems. 2013 Nov 1;38(8):1070-83.

(25) Tang X, Xu A. Multi-class classification using kernel density estimation on K-nearest neighbours. Electronics Letters. 2016 Mar 24;52(8):600-2.

(26) Benites F, Sapozhnikova E. Improving scalability of ART neural networks. Neurocomputing. 2017 Mar 22;230:219-29.

(27) Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K. A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919. 2017 Jul 10.

(28) Aliwy AH, Ameer EA. Comparative study of five text classification algorithms with their improvements. International Journal of Applied Engineering Research. 2017;12(14):4309-19.

(29) Tsangaratos P, Ilia I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena. 2016 Oct 1;145:164-79.

(30) Calma, Adrian, Tobias Reitmaier, and Bernhard Sick. Semi-supervised active learning

for support vector machines: A novel approach that exploits structure information in data. Information Sciences 456 (2018): 13-33.

(31) Zhu XJ. Semi-supervised learning literature survey. University of Wisconsin-Madison Department of Computer Sciences; 2005.

(32) Korde, Vandana, and C. Namrata Mahender. Text classification and classifiers: A survey. International Journal of Artificial Intelligence Applications 3, no. 2 (2012): 85.

(33) Dwivedi SK, Arya C. Automatic text classification in information retrieval: A survey. InProceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies 2016 Mar 4 (p. 131). ACM.

(34) Shafiabady N, Lee LH, Rajkumar R, Kallimani VP, Akram NA, Isa D. Using unsupervised clustering approach to train the Support Vector Machine for text classification. Neurocomputing. 2016 Oct 26;211:4-10.

(35) Liu, Haijun, Jian Cheng, and Feng Wang. Sequential subspace clustering via temporal smoothness for sequential data segmentation. IEEE Transactions on Image Processing 27, no. 2 (2017): 866-878.

(36) Liu, Bing. Opinion mining and sentiment analysis. In Web Data Mining, pp. 459-526. Springer, Berlin, Heidelberg, 2011.

(37) Liu, Yuchen, Xiaochuan Ni, Jian-Tao Sun, and Zheng Chen. Unsupervised transactional query classification based on webpage form understanding. In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 57-66. ACM, 2011.

(38) Yan, W., Zhang, B., Ma, S., Yang, Z. (2017). A novel regularized concept factorization for document cluster- ing. Knowledge-Based Systems, 135(1), 147-158. https://doi.org/10.1016/j.knosys.2017.08.010.

(39) Ahmad S, Lavin A, Purdy S, Agha Z. Unsupervised real-time anomaly detection for streaming data. Neurocomputing. 2017 Nov 1;262:134-47.

(40) Shi, Lei, Xinming Ma, Lei Xi, Qiguo Duan, and Jingying Zhao. Rough set and ensemble learning based semi-supervised algorithm for text classification. Expert Systems with Applications 38, no. 5 (2011): 6300-6306.

(41) Santos, A., Canuto, A. (2014). Applying semi-supervised learning in hierarchical multi-label classification. Expert Systems with Applications, 41, 6075–6085. https://doi.org/10.1016/j.eswa.2014.03.052.

(42) Park J. Simultaneous estimation based on empirical likelihood and general maximum likelihood estimation. Computational Statistics Data Analysis. 2018 Jan 1;117:19-31.

(43) Brownlee J. Supervised and Unsupervised Machine Learning Algorithms. Retrieved March 11, 2018.

(44) Albitar, Shereen, Bernard Espinasse, and Sébastien Fournier. Towards a Supervised Rocchio-based Semantic Classification of Web Pages. In KES, pp. 460-469. 2012.

(45) Partalas, Ioannis, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artieres, George Paliouras, Eric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, and Patrick Galinari. Lshtc: A benchmark for large-scale text classification. arXiv preprint arXiv:1503.08581 (2015).

(46) Maalouf M, Siddiqi M. Weighted logistic regression for large-scale imbalanced and rare events data. Knowledge-Based Systems. 2014 Mar 1;59:142-8.

(47) Yen SJ, Lee YS, Ying JC, Wu YC. A logistic regression-based smoothing method for Chinese text categorization. Expert Systems with Applications. 2011 Sep 1;38(9):11581-90.

(48) Aseervatham S, Antoniadis A, Gaussier É, Burlet M, Denneulin Y. A sparse version of the ridge logistic regression for large-scale text categorization. Pattern Recognition Letters. 2011 Jan 15;32(2):101-6.

(49) An Y, Tang X, Xie B. Sentiment analysis for short Chinese text based on character-level methods. In2017 9th International Conference on Knowledge and Smart Technology (KST) 2017 Feb 1 (pp. 78-82). IEEE.

(50) Pereira JM, Basto M, da Silva AF. The logistic lasso and ridge regression in predicting corporate failure. Procedia Economics and Finance. 2016 Jan 1;39:634-41.

(51) Sucar, L.E. (2015). Bayesian classifiers. In L. E. Sucar, Probabilistic graphical models (pp. 41-62). Springer. https://doi.org/10.1007/978-1-4471-6699-3-4.

(52) Jiang, Liangxiao, Chaoqun Li, Shasha Wang, and Lungan Zhang. Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence 52 (2016): 26-39.

(53) Diab DM, El Hindi KM. Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. Applied Soft Computing. 2017 May 1;54:183-99.

(54) Arar ÖF, Ayan K. A feature dependent Naive Bayes approach and its application to the software defect prediction problem. Applied Soft Computing. 2017 Oct 1;59:197-209.

(55) LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature 521, no. 7553 (2015): 436-444.

(56) Cerri, R., Barros, R. C., Carvalho, A. (2014). Hierarchical multi-label classification using local neural net- works. Journal of Computer and System Sciences, 80, 39–56. https://doi.org/10.1016/j.jcss.2013.03.007.

(57) Nie Q, Jin L, Fei S, Ma J. Neural network for multi-class classification by boosting composite stumps. Neurocomputing. 2015 Feb 3;149:949-56.

(58) Demidova L, Klyueva I, Sokolova Y, Stepanov N, Tyart N. Intellectual approaches to improvement of the classification decisions quality on the base of the SVM classifier. Procedia Computer Science. 2017 Jan 1;103:222-30.

(59) Manevitz, Larry M., and Malik Yousef. One-class SVMs for document classification. Journal of machine Learning research 2, no. Dec (2001): 139-154.

(60) Sun, Aixin, and Ee-Peng Lim. Hierarchical text classification and evaluation. In Proceedings 2001 IEEE International Conference on Data Mining, pp. 521-528. IEEE, 2001.

(61) Altınel, Berna, Murat Can Ganiz, and Banu Diri.

(62) Goudjil M, Koudil M, Bedda M, Ghoggali N. A novel active learning method using SVM for text classification. International Journal of Automation and Computing. 2018 Jun 1;15(3):290-8.

(63) Ramesh B, Sathiaseelan JG. An advanced multi class instance selection based support vector machine for text classification. Procedia Computer Science. 2015 Jan 1;57:1124-30.

(64) Tbarki K, Said SB, Ksantini R, Lachiri Z. One-class SVM for landmine detection and discrimination. In2017 International Conference on Control, Automation and Diagnosis (ICCAD) 2017 Jan 19 (pp. 309-313). IEEE.

(65) Gupta V. Recent trends in text classification techniques. InInternational Journal of Computer Applications 2011 (Vol. 35, No. 6, pp. 45-51). International Journal of Computer Applications, 244 5 th Avenue, 1526, New York, NY 10001, USA India.

(66) Maillo, J., Ramfrez, S., Triguero, I., Herrera, F. (2016). kNN-IS: An iterative spark-based design of the k- nearest neighbors classifier for big data. Knowledge-Based Systems, 117, 3-15. https://doi.org/10.1016/j.knosys.2016.06.012.

(67) Liu, Xuejie, Jingbin Wang, Ming Yin, Benjamin Edwards, and Peijuan Xu. Supervised learning of sparse context reconstruction coefficients for data representation and classification. Neural computing and applications 28, no. 1 (2017): 135-143.

(68) Han, Eui-Hong Sam, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In Pacific-asia conference on knowledge discovery and data mining, pp. 53-65. Springer, Berlin, Heidelberg, 2001.

(69) Sahgal, Divya, and A. Ramesh. On Road Vehicle Detection Using Gabor Wavelet Features with Various Classification Techniques. In Proceedings of the 14th International Conference on Digital Signal Processing Proceedings. DSP. 2002.

(70) Kotsiantis SB. Decision trees: a recent overview. Artificial Intelligence Review. 2013 Apr 1;39(4):261-83.

(71) Savas, S. K., Nasibov, E. (2017). Fuzzy ID3 algorithm on linguistic dataset by using WABL defuzzification method. Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples. Italy. https://doi.org/10.1109/FUZZ-IEEE.2017.8015502.

(72) Nguyen, Thanh-Tung, Huong Nguyen, Yinxu Wu, and Mark Junjie Li. Classifying gene data with regularized ensemble trees. In 2015 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 1, pp. 134-139. IEEE, 2015..

(73) Sun, Z., Ye, Y., Deng, W., Huang, Z. (2011). A cluster tree method for text categorization. Procedia Engineering, 15, 3785-3790. https://doi.org/10.1016/j.proeng.2011.08.709.

(74) Al Zamil MG, Can AB. ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. Knowledge-Based Systems. 2011 Feb 1;24(1):58-65.

(75) Asadi S, Shahrabi J. Complexity-based parallel rule induction for multiclass classification. Information Sciences. 2017 Feb 20;380:53-73.

(76) Vieira AS, Borrajo L, Iglesias EL. Improving the text classification using clustering and a novel HMM to reduce the dimensionality. Computer methods and programs in biomedicine. 2016 Nov 1;136:119-30.

(77) Gomez JC, Boiy E, Moens MF. Highly discriminative statistical features for email classification. Knowledge and information systems. 2012 Apr 1;31(1):23-53.

(78) Cerchiello P, Giudici P. Non parametric statistical models for on-line text classification. Advances in Data Analysis and Classification. 2012 Dec 1;6(4):277-88.

(79) Reitmaier, Tobias, Adrian Calma, and Bernhard Sick. Semi-supervised active learning for support vector machines: A novel approach that exploits structure information in data. arXiv preprint arXiv:1610.03995 (2016).

(80) Mirzamomen Z, Kangavari MR. Evolving fuzzy min–max neural network based decision trees for data stream classification. Neural Processing Letters. 2017 Feb 1;45(1):341-63.

(81) Karabadji NE, Seridi H, Bousetouane F, Dhifli W, Aridhi S. An evolutionary scheme for decision tree construction. Knowledge-Based Systems. 2017 Mar 1;119:166-77.

(82) Hiew BY, Tan SC, Lim WS. Intra-specific competitive co-evolutionary artificial neural network for data classification. Neurocomputing. 2016 Apr 12;185:220-30.