Modified Algorithm for Mining Medical Terms having Multi-Label using Hybrid Classifier

[1]Mamta Patel,[2]Kinjal Thakar
[1]M.E. Student,[2]Assistant Professor
[1]Department of Information Technology,
[1]Silver Oak College of Engineering & Technology, Ahmedabad, India

Abstract : Text mining technology is widely applied to a various fields. Semantic web, Information retrieval, Natural language processing and Machine learning text mining are influential areas in text mining. Multi-label learning has obtained important attention in the research area from previous few years. Medical data mining has large prospective for exploring the unrevealed patterns in the data sets of the medical field. Medical diagnosis plays significant role but still complex task that needs to be performed precisely and effectively. Applying automation of this system will be beneficial in medical science. Mining medical records for relationships between alive elements and the symptoms of a disease is an significant task. In this paper, we modify algorithm for mining medical terms having multi-label. This Research Paper leads to Decision Support System, Improved Health, Personalized Medicine and etc.

IndexTerms **- Machine Learning, NLP, LDA, Hybrid Classifier.**

I. INTRODUCTION

Text mining, which is here and there referred as "text analytics" is one approach to make subjective or "unstructured" data usable by a PC.

Qualitative text values are illustrative information that can't be estimated in numbers and regularly incorporates characteristics of appearance like shading, surface, and literary portrayal. Quantitative text is numerical, organized information that can be estimated. Be that as it may, there is regularly slippage among qualitative and quantitative classes. For instance, a photo may generally be considered "subjective information" however when you separate it to the dimension of pixels, which can be estimated. Text mining technology is widely applied to various fields like e-business, Marketing and Health care. Text mining is the task of extracting meaningful information from large textual database. Text mining is similar to data mining except that data mining can only handle structured data sets but text mining can handle unstructured or semi-structured data set like emails, HTML files etc. So text mining is better for handle large volume of data[8]. Multi-label learning has obtained important attention in the research area from previous few years. Extremely large data is available in medical field, using this data we can diagnosis many diseases by text mining techniques in effective manner. Medical document contain important information for diagnosing diseases Such as combination of signs, symptoms, and test results used to determine the correct diagnosis. Multi-label classification was mostly used in the medical diagnosis task. This Research paper states Informed Decisions, Probability measures, Predictive modeling, Improved Health and personalized Medicine.

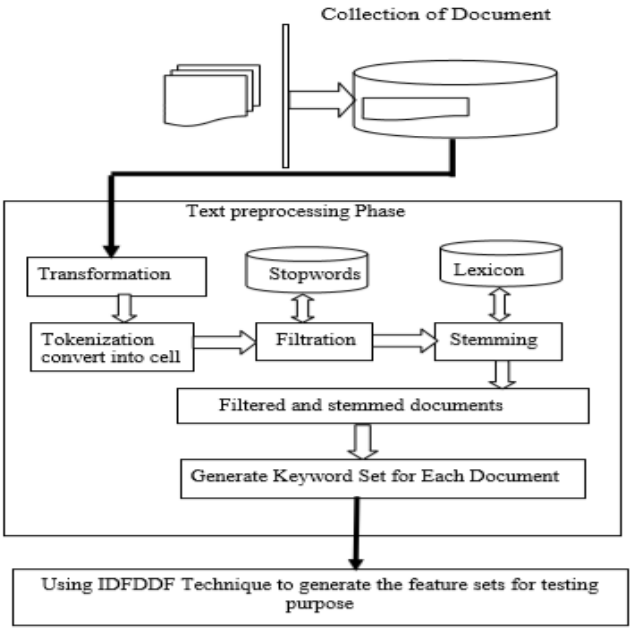

Figure 1. Text Mining Process for Classification Results

There can be other feature selection and feature extraction techniques and algorithms exist and as they are popular among programmers all over the world, these methods are designed in such a way that to remove duplicate and irrelevant features so that classification of new instances will be more accurate. The whole process is divided in stages as stated in this diagram.Machine learning algorithms tend to be affected by noisy data. noise ought to be decreased however much as could reasonably be expected with the end goal to maintain a efficiency and proficiency of the algorithm in the indirect models and improve the efficiency of the algorithm. LDA tries to project the sample onto a straight line so that the projection points of interclass samples are as close as possible and the projection points of the intra class samples are as far apart as possible. When classifying a new sample, we projected it onto the same line[18]. In this paper, experiments are conducted on medical dataset of clinical free text reports from MEDLINE Data [17].

II. RELEVANT WORK

In [1] paper, utilization of highlight choice strategies for enhancing multi-mark clinical text order is talked about. We will join issue change strategies with various ways to deal with highlighted chosen algorithms. This paper has utilized Select K Best (SKB) technique with clearly understood $\chi 2$ (chi-square) measurement strategy [1]. There were analyzed filter and wrapper techniques together with mixture approaches. The execution of the considered methodologies is looked at by utilizing two measurements: Classification Accuracy and Hamming Loss, thinking about various number of marks and information occurrences. Tests were performed on the course of action of clinical reports and exhibited the advantage of the channel technique and promising results for the cream system. Constraint of this paper is including name conditions into highlight choice is still to be finished. There is a lot of useful information available in medical documents. Information as medical named entities, relationship between medical entities, medical summary and etc.

Most of the time such information in medical documents is unstructured and available in nonstandard natural language so it is difficult to automatically collect and present this information in a structured way [2]. In [2] this article we propose a feature based machine learning model for relation classification task. There are plenty of medical documents available and it is a complex task to extract the information from these clinical documents. Relation Extraction task is to find whether a pair of medical entities is related or not. Clinical entities are the named entities in the medical domain such as Treatment, Problems, and Procedure etc. Our task is to get the specific relation between a given pair of medical entities. Improve the feature based relation classification method using a richer and important feature set. train model using SVM classifier with many such examples from i2b2 (2010) clinical data set. observed such features after lots of study on medical text and included those features in the existing relation classification model and got the better result.

In [3] Motivated by the ZHENG classification, this paper has proposed a novel segmentation method. This paper has also machine learning methods which were used to evaluate four feature extraction methods' capability in extracting useful information for psoriasis disease with ZHENG classification from medical texts. As psoriasis is interactable and its cause is difficult to discover, Traditional Chinese Medicine(TCM) is proved in China to be a more effective medical way in treating psoriasis. TCM has its unique techniques for diagnosis and treatment. "Treatment based on ZHENG differentiation" is one of the essential theory. ZHENG, which means a characteristic profile of all clinical manifestations identified by a TCM practitioner, is regarded as the target in the diagnostic process instead of disease. And decision on prescription is based on ZHENG rather than disease. this paper presents our experiment results of various ZHENG classification models with different feature extraction method using dif ferent machine learning methods, and then provide suggestions about the feature extraction usage. To achieve goal, bow and word2vec were used respectively to extract features from text data. extracted factor sets as features to train models and evaluate the models' accuracy, precision, recall and F-measure. techniques of ZHENG identification can be learned only by the one or two apprentices of a particular TCM expert, which is quite inefficiently. Multi label based Chinese text categorization is achieved by evolution of POCA and WOCA algorithm.

This paper[4], has proposed a multi-label learning algorithm for Chinese text categorization. Multi-label learning problem is transformed into traditional binary classification problem. In this paper, two kinds of training datasets construction methods: Position-in-Order Construction Algorithm (POCA) and Weight-in- Order Construction Algorithm (WOCA). Both methods can avoid the combination explosion problem of training datasets and ensure that each label of a sample can appear in one of the ensemble training datasets. There are two methods to derive the single-label multi-class datasets from original multi-label dataset to train the SVM-based binary classifiers. we test the performance of multi-label classifier with Subset accuracy, Hamming loss, Micro-average and Macro-average. This is an extremely challenging problem because of the particularity of Chinese grammar.
Classification learning methods, such as decision tree, SVM, BNN or deep learning, are applied for construction of classifiers as mentioned in this paper[5]. Experimental validation shows that random forest achieved the best performance and the second best was the deep learner with a small difference , but decision tree methods with many keywords performed only a little worse than neural network or deep learning methods. Different approaches are proven apt for yielding useful results on variety of text and contents.

In [6] this paper, introduced work grouping of multi-space records is performed by utilizing Weka-Lib SVM classifier. Here to change gathered preparing set and test set reports into term-archive network (TDM), the vector space demonstrate is utilized. In classifier TDM is utilized to create anticipated outcomes. The outcomes rose up out of Weka with its GUI support utilizing TDM have snappy reaction time in arranging the archives. At its least difficult, framework gives a brisk and simple approach to

investigate and break down information. What's more the reaction time for arranging archives gets definitely lessened. Impediment of this paper is available system still to be improved.

In [7] this paper, a new feature selection method called Optimized Swarm Search-based Feature Selection (OS-FS) is proposed. The swarm search in OS-FS is optimized by a new feature evaluation technique called Clustering-by-Coefficient-of-Variation (CCV). CCV is one of the fastest approach in finding appropriate features for classification. The results show superiority of OS-FS over the traditional feature selection methods. Often the textual words are not used directly in the model induction. During preprocessing the words in a document are converted to word vectors. Depending on the transformation methods, usually it counts the frequencies of occurrence of the words present in the document. The extracted texts are converted into sparse matrix by using a Weka Filter namely *StringToWordVector*. CCV was designed in mind that operates fast and accurate, as an alternative feature selection approach. It is grounded on the principle of Standardized Measure of Dispersion (SMD). SMD has the advantage that the coefficients of dispersion are independent of the units of observation. The results unanimously show that OS-FS can improve the default swarm-based FS in Weka with certain gain in both accuracy and Kappa. Disadvantage of OS-FS is swarm search usually requires longer run-time because of its iterative characteristic.

III. MOTIVATION

The objective of all studied research is better classification and filtration of text in more efficient way. All the experiments are performed based on different classification methods which are naive bayes or SVM. But label dependencies are still required to be proposed in such research types [1]. Methods used to train the model can still be improved in paper [2]. Techniques of ZHENG identification can be learned only by a particular TCM expert, which is quite inefficiently [3]. Multi-label Learning Algorithm for Chinese Text Classification [4] is an extremely challenging problem because of the particularity of Chinese grammar. Medical terms are not standardized in term of terminology in data mining researches. Therefore variety of terminology has been added. Some words are added based on medical vocabulary and some are added based on data analysis process. That is the reason which makes them difficult to analyze and extract patterns [5]. There is a limitation of paper [6] present technique to be enhanced and implemented in real time in firewall for allowing the documents of particular needed domain if validated rather than blocking the anonymous sites. Feature selection has long been a problem in data mining which can be considered as finding the right feature subset in a huge state-space search when the data are highly dimensional. swarm search usually requires longer run-time because of its iterative characteristic [7].

IV. PROPOSED WORK

In this section, method used for our proposed work is discussed in detail.

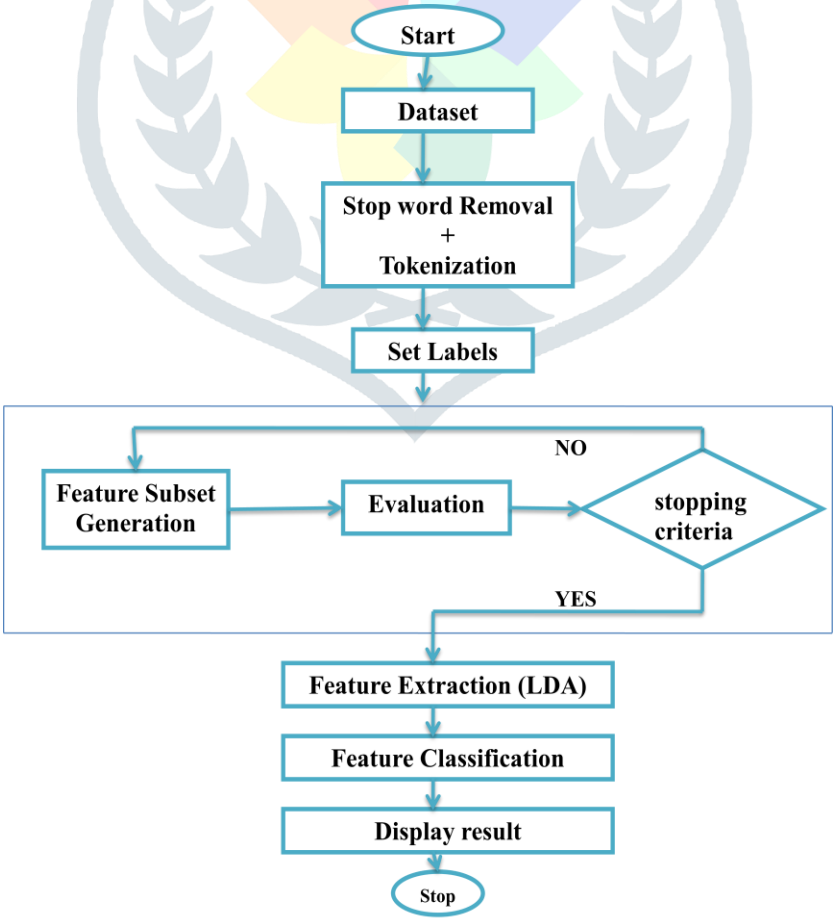

Figure 2. Work Flow of Proposed Algorithm

Proposed flow work for dataset as input. The specified dataset is used to test algorithm. next Process will be performed for document pre processing. Data pre-processing is an often neglected but important step in the data mining process because If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. The Pre-processing Steps such as Stop word and tokenization should be performed to set labels to documents. After labels are set, Perform feature selection. Feature selection is used because it removes irrelevant, redundant and noisy data from dataset. A feature selection process primarily consists of three steps namely subset generation, subset evaluation and stopping criterion. Subset generation generates a feature subset which is the candidate for evaluation. The generated subset is then evaluated by using an evaluation criterion which determines the worthiness of the subset. The subset generated is then compared with the best subset generated earlier. Based on a stopping criterion, further process of generation and evaluation of subsets is not continued. Then Perform Feature Extraction using Linear Discriminant analysis(LDA). Extract meaningful information from original data is referred as a feature extraction. Feature extraction results in a much smaller and richer set of attributes. LDA measures the relevance of features by their correlation with dependent variables. with LDA we need to find a new feature space to project the data in order to maximize classes seperability. now, we have to perform feature classification on relevance data obtained by feature extraction. Feature classification is the grouping of features based on some criteria. LDA is used for calculate distance between labels. Naive Bayes Classifier and Multi-label KNN(ML-KNN) are used in this approach. Naive Bayes is used for prediction of labels based on relevance data obtain by LDA. ML-KNN is an extension of the KNN for multi-label data.ML-KNN classifier display result on MEKA toolkit.

Algorithm

Input: Dataset
Output: After applying algorithm medical terms obtained.

1. Read document and pre-process stop word removal and tokenization done, labels are set to document.
2. Perform Label wise Feature Selection.
3. Generation of subset of features and evaluate it.
4. **if** there is stopping criteria **then** perform feature extraction using linear discriminate allocation(LDA) Calculated as

   $$I(w) = \frac{w^T(\sum(x_i-\mu)^T(x_i-\mu)w)}{w^T(\sum_c\sum_{iec}(x_i-\mu_c)^T(x_i-\mu_c)w))}$$

   else

   go to 3
5. Feature classification method Naive Bayes(NB) and Multi-Label KNN(ML-KNN) are applying to Data Calculated as

   $$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

   Where P(A|B) is "Probability of A given B",
   P(A) is Probability of A
   P(B|A) is "Probability of B given A"
   P(B) is Probability of B

V. RESULTS

This paper presents semantic based study of medical text based on different diseases and record. The information is obtained from reliable sources like government approved websites, kaggle and github. The main aim of medical text mining is to use NLP for analyzing medical terms but not just English or any other linguistic words. The another aim is classifying, recognizing and obtaining labels of the given data. In this paper, we proposed a hybrid algorithm and performed text analysis of medical data. The classification and LDA proves to be simple, versatile and feasible as compared to other machine learning algorithms. Median Of Categorization Results For All Classifiers And Feature Selection Methods After Br And Lp Data Transformations, Classification Accuracy (Ca) [%].

TABLE I
CLASSIFICATION ACCURACY (CA) [%]

		2 labels for instance		
	baseline	BR trans.	LP trans.	LP ↑[a]
BR	33.27	38.95	41.81	2.86
LP	46.93	**55.17**	**55.85**	0.67
CC	37.75	41.54	45.33	**3.79**
ECC	38.68	41.76	45.52	3.77

LC	**46.94**	54.96	55.19	0.23
PA	**50.00**	**52.00**	**55.90**	**3.632**

TABLE II
CLASSIFICATION ACCURACY (CA) [%]

		3 labels for instance	
	baseline	BR trans.	LP trans.
BR	2.50	15.00	15.00
LP	15.00	20.00	20.00
CC	0.00	10.00	10.00
ECC	7.50	10.00	10.00
LC	**25.00**	**28.57**	**28.57**
PA	**25.05**	**28.62**	**28.62**

TABLE III
CLASSIFICATION ACCURACY (CA) [%]

		All dataset (2 and 3 labels)		
	baseline	BR trans.	LP trans.	LP $\uparrow^a$
BR	31.83	36.69	39.37	2.68
LP	44.95	53.01	53.75	0.75
CC	36.03	39.12	43.00	**3.88**
ECC	37.06	40.10	43.72	3.62
LC	**45.36**	53.98	**54.20**	0.22
PA	**47.80**	**53.05**	**54.25**	**3.90**

Classification improvement after LP data transformation (compared to the BR transformation),where PA is Proposed approach)

The above table I,II & III show the classification and identification process for n_label and allow_labeled with multi label classifiers. Here classes c are the classes (neonatal,skin,gastric,blood,spinal,neuro,catheter,hypotensive,infectious, terbinafine). Here, xjxj is the value vector per instance per class (we have in our case two dimensions x and y, so for instance x1x1 has the dimensionality 2x1). μcμc represents the mean-vector of class cc and is a vector which contains the values of each dimension for each class.

TABLE IV
HAMMING LOSS (HL) [%]

		2 labels for instance		
	baseline	BR trans.	LP trans.	LP $\downarrow^a$
BR	3.50	**2.36**	**2.28**	-0.07
LP	3.27	2.56	2.50	-0.06
CC	3.38	2.67	2.64	-0.03
ECC	3.24	2.61	2.62	0.01
LC	**3.21**	2.41	2.33	**-0.09**
PA	**3.09**	**2.00**	**2.10**	**-0.10**

The above tables are demonstrating results of proposed approach in comparison with the state of the art techniques of multi label classifier. The classification and identification process is performed for 1000 instances of class1 to class10.

TABLE V
HAMMING LOSS (HL) [%]

		3 labels for instance		
	baseline	BR trans .	LP trans.	LP $\downarrow^a$
BR	6.56	4.67	4.67	-
LP	5.45	5.56	5.56	-
CC	6.39	4.78	4.67	**-0.11**

ECC	5.95	4.73	4.67	-0.05
LC	**4.29**	**3.89**	**3.89**	-
PA	**4.21**	**2.38**	**2.99**	**-0.11**

The above table V is demonstrating results of proposed approach that uses Binary relevance & Classifier Chain with label 3 scenarios. The transactions analysis with Label power set & Label chain are prediction base results. The classification and identification process is performed for 1000 instances of class1 to class3.

TABLE VI
HAMMING LOSS (HL) [%]

		All dataset (2 and 3 labels)		
	baseline	BR trans.	LP trans.	LP ↓[a]
BR	3.74	2.51	2.43	**-0.08**
LP	3.47	2.74	2.70	-0.04
CC	372	2.79	2.75	-0.04
ECC	3.59	2.74	2.75	0.01
LC	**3.52**	**2.43**	**2.41**	-0.02
PA	**3.17**	**2.09**	**1.983**	**-0.092**

Classification improvement after LP data transformation (compared to the BR transformation),where PA is Proposed approach

Final results in Table VI displays combined average results for all type of class labels. Here it shows the proposed approach find approx 10% better results in comparison with LC and for all classifier.

TABLE VII

Results for different document sets			
	precision	**recall**	**Accuracy**
DOC-1	78.02	84.05	80.92
DOC-2	74.05	78.00	75.97
DOC-3	87.25	87.70	87.47
DOC-4	85.05	86.55	85.79
DOC-5	76.35	78.54	77.42
DOC-6	80.25	85.35	82.72

The table VII is demonstrating the results of precision, recall and accuracy for different documents sets selected to test multi label classification. Each documents consists of texual information about a medical case diagnosis. The formula used for accuracy is f=2*(precision * recall) / (precision + recall), and precision is calculated based on standard deviation of highest and lowest value found from the results.

VI. CONCLUSION

Medical data records are analyzed before further processing of the proposed approach. The analysis results have shown that the content are in varied size and types. Many times they are less textual and more pictorial. Feature subset generation is the process to make the contents into a process able data. Here, Research were conducted on medical text reports displayed the comparison of the base paper method and proposed approach. We are expecting to get good result of the new methodology as its displayed in the table I & II. In this paper, application of the feature selection technique which will be applied not only single label approach but also on multi-label approach. Here we use LDA as a feature selection method. Most feature selection methods concerns filter methods.

In future, we need to put more effort in increasing the performance of the algorithm. Challenges such as the presence of sarcasm, blind negation, complex word meanings, spam detection, forged data, fuzzy data, handling hidden features could be taken up as research areas.

REFERENCES

[1] Kinga Glinka, Rafał Wozniak and Danuta Zakrzewska,"Improving Multi-Label Medical Text Classification by Feature Selection",IEEE,2017.

[2] Saumaya Gupta and Amit Kumar Manjhvar," Relation Classification from Unstructured Medical Text using Feature Based machine Learning Approach",IEEE,2017.

[3] Zehui He, Heng Weng, Aihua Ou, Shixing Yan, Chuanjian Lu and Guo-Zheng Li," Feature Extraction from Medical Record Text for TCM Zheng Classification of Psoriasis", IEEE,2017.

[4] Xun Wang, Huan Liu, Zeqing Yang, Jiahong Chu, Lan Yao, ZhiBin Zhao, Bill Zuo," Research and Implementation of a Multi-label Learning Algorithm for Chinese Text Classification", IEEE,2017.

[5] Shusaku Tsumoto, Tomohiro Kimura, Haruko Iwata, and Shoji Hirano,"Mining Text for Disease Diagnosis in Hospital Information System",IEEE,2017.

[6] Uma Somani, Kanika Lakhani, Manish Mundra ," A Novel Data Mining Approach for Multi Variant Text Classification",IEEE,2015.

[7] Simon Fong, Elisa Gao and Raymond Wong," Optimized Swarm Search-based Feature Selection for Text Mining in Sentiment Analysis",IEEE,2015.

[8] Revathi M Nair and Sindhu L," A Survey on Medical Text Mining", International Journal of Computer Applications, December 2014.

[9] N. Spolaôr, M. C. Monard, G. Tsoumakas, and H. D. Lee, "A systematic review of multi-label feature selection and a new method based on label construction," *Neurocomputing*, vol. 180, pp. 3–15, 2016.

[10] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, "Categorizing feature selection methods for multi-label classification," *Artificial Intelligence Review*, pp. 1–22, 2016.

[11] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004

[12] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelec- tronics (MIPRO)*, May 2015, pp. 1200–1205.

[13] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[14] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelec- tronics (MIPRO)*, May 2015, pp. 1200–1205.

[15] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[16] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in *Feature extraction*. Springer, 2006, pp. 137–165.

[17] "MEDLINE Data." [Online]. Available: https://www.nlm.nih.gov/databases/download/pubmed_medline.html

[18] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4480804/

[19] P. Refaeilzadeh, L. Tang, and H. Liu, "On comparison of feature selection algorithms," in *Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II*, 2007, pp. 34–39.