Australasian Physical & Engineering Sciences in Medicine (2018) 41:1087–1100
https://doi.org/10.1007/s13246-018-0674-3

SPECIAL ISSUE ARTICLE

Medical data mining in sentiment analysis based on optimized swarm search feature selection

Daohui Zeng[1] · Jidong Peng[2] · Simon Fong[3] · Yining Qiu[4] · Raymond Wong[4]

Received: 28 June 2018 / Accepted: 9 August 2018 / Published online: 11 September 2018

Abstract

In this paper, we propose a novel technique termed as optimized swarm search-based feature selection (OS-FS), which is a swarm-type of searching function that selects an ideal subset of features for enhanced classification accuracy. In terms of gaining insights from unstructured medical based texts, sentiment prediction is becoming an increasingly crucial machine learning technique. In fact, due to its robustness and accuracy, it recently gained popularity in the medical industries. Medical text mining is well known as a fundamental data analytic for sentiment prediction. To form a high-dimensional sparse matrix, a popular preprocessing step in text mining is employed to transform medical text strings to word vectors. However, such a sparse matrix poses problems to the induction of accurate sentiment prediction model. The swarm search in our proposed OS-FS can be optimized by a new feature evaluation technique called clustering-by-coefficient-of-variation. In order to find a subset of features from all the original features from the sparse matrix, this type of feature selection has been a commonly utilized dimensionality reduction technique, and has the capability to improve accuracy of the prediction model. We implement this method based on a case scenario where 279 medical articles related to 'meaningful use functionalities on health care quality, safety, and efficiency' from a systematic review of previous medical IT literature. For this medical text mining, a multi-class of sentiments, positive, mixed-positive, neutral and negative is recognized from the document contents. Our experimental results demonstrate the superiority of OS-FS over traditional feature selection methods in literature.

Keywords Medical text mining · Optimized swarm search-based feature selection · Sentiment prediction · Clustering-by-coefficient-of-variation

Introduction

Based on the context and nature of medical-based articles, sentiment prediction is an important research topic that provides indications of a large volume of medical texts to be abstracted into emotions. Journal articles which are in the format of unstructured text, either digitalized or in hard-copies continue as the primary media for publishing biomedical research results. For example, the MEDLINE database of 5639 selected publications covering biomedicine and health from 1950 to 2013,[1] greater than 21.6 million records can

Daohui Zeng and Jidong Peng contributed equally to this work and are co-first authors.

✉ Simon Fong
ccfong@umac.mo

Daohui Zeng
zengdh1971@163.com

Jidong Peng
pengjidong66@sina.com

Yining Qiu
yining899926@hotmail.com

Raymond Wong
wong@cse.unsw.edu.au

[1] https://www.nlm.nih.gov/bsd/num_titles.html.

1 First Affiliated Hospital of Guangzhou University of TCM, Guangzhou, People's Republic of China

2 Ganzhou People's Hospital, Jiangxi, People's Republic of China

3 Department of Computer and Information Science, University of Macau, Taipa, Macau SAR, People's Republic of China

4 School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

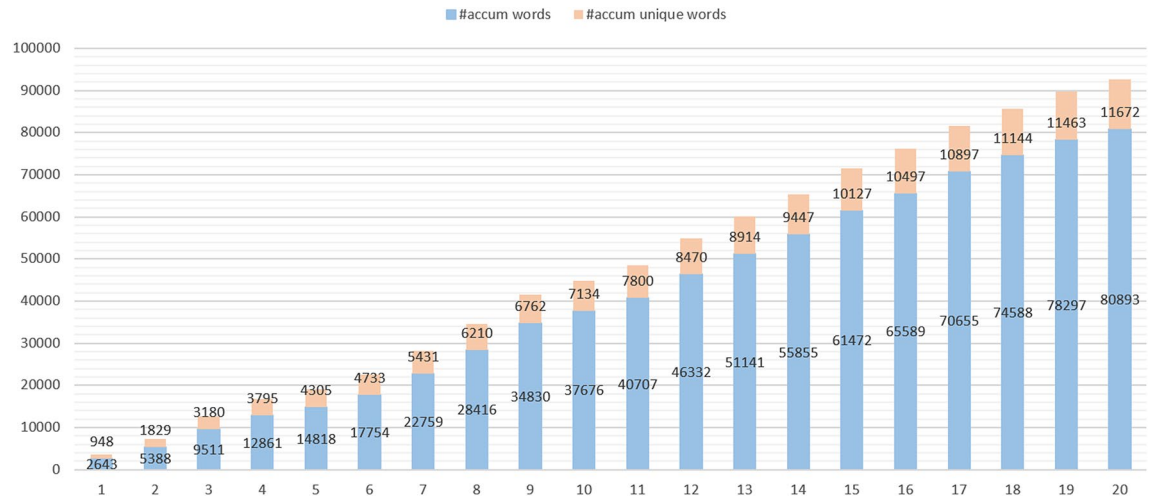

Fig. 1 The accumulated unique word counts versus the number of articles in the training dataset

be assessed by PubMed. New references and abstracts on life sciences and biomedical topics are added at an increasing rate of more than half a million each year. Automated tools such as those advanced from supervised machine learning and natural languages have been deployed to extract, to understand and to predict the readers' emotions towards such large volume of articles.

To achieve sentiment prediction based on the fundamental data analytics, medical text mining is often implemented. Here, a popular choice in text mining for sentiment prediction is the classification technique. The non-deciphered texts extracted from social media or other online document sources are pre-processed into computerized format. Note that textual words are not used directly in the model induction for the underlying mappings between the target classes and the words or their binary representatives. In the sentiment prediction framework, the conversion of words to the word vectors in the medical document is performed during pre-processing stage. Typically, the strategy is to count the frequencies of occurrence of the words in the document, and this is dependent on the transformation techniques available. The text transformation output is a sparse matrix of binary Booleans with dimension as large as the total number of unique words or phrases in the document. Each row is a record of document formatted as a bitwise vector, called word vector representing the words that exist in the text of the document. Based on the collection of all the words that could be found from the set of training documents, knowing which words that pertain to an individual document is required; and a classifier is then able to compute all mapping relations between those words that characterize the document and a particular target class. But note that the textual words first need to be transformed into bit vectors that are to be stored in memory during model construction prior to the phase of model construction.

Aims and objectives

The sparse matrix is usually huge in terms of size despite the fact that text transformation is effective by its simple principle of word frequencies in text mining. When the training dataset is unbounded, and the data input is continuous such as data stream or live data feeds, this type of problems tend to occur frequently. In the internet environment, Tweets from different users tends to pile up over time, after comments on health-care and medical treatments are added at various forum, and this type of data feeds can be approaching infinity. The computational challenge in run-time memory and effective induction for an accurate classifier comes about when we transform the texts into one huge sparse matrix. For example, the increase of the sparse matrix size in terms of the number of dimensions with the quantity of information that are selectively extracted based on selected articles of a journal called "Health Information Technology: An Updated Systematic Review with a focus on Meaningful Use Functionalities" published by Healthit.gov (Fig. 1).

Based on the Weka Filter namely *StringToWordVector*,[2] extracted texts can be converted into sparse matrix. Here, Weka is a Java based open source machine learning benchmarking platform (University of Waikato, NZ). The filter transforms text strings by tokenizing the words from the training document into a set of binary features representing word occurrence. A sparse matrix has the columns/attributes

[2] http://wiki.pentaho.com/display/DATAMINING/StringToWordVector.

that indicate whether or not the particular word is labeled as the attribute, and they are presented in the rows of text entries.

Dependent on the unique words that are contained in the input documents, the features set is converted. At different degrees linearly, the filter's buffered and non-buffered modes results in high dimension of sparse matrix, which pertains to increase rates of 18% and 12% for the non-buffered and buffered modes respectively.

In terms of subsiding the curse of dimensionality problem, a good option may be the feature selection. The input texts are daily news and they may potentially amount to infinity given the sentiment prediction scenario. Effective and efficient dimensionality reduction is required as dimension would be unconstrained. Here, effective feature selection means the algorithm is able to choose a best or an almost best feature subset; whereas efficiency means the algorithm has to do so within a reasonable time instead of brute-force, assuming the need to support real-time online analytics.

Advantage of optimized swarm search-based feature selection (OS-FS)

The stochastic-based feature selection method called optimized swarm search-based feature selection (OS-FS) is proposed in lieu of other sequential search methods. OS-FS is a swarm-type of searching function that selects an ideal subset of features for enhanced accuracy in classification. Here, there is no need to process through the whole spectrum of dimensions of the sparse matrix using swarm search. In this technique, selected dimensions are sampled into combinations and tested for suitability by probabilistic movement. One disadvantage of swarm search is the lack of guarantee of a perfect solution because it operates by probability, and it is not meant to be a solution that covers all possibilities. In addition, swarm search does not necessarily have to process through all combinations of subsets from the given dimensions of the sparse matrix. The swarm search obtains an optimal solution between time limit and highest possible accuracy without a throughout search of all the dimensions that may take an extremely long time.

Now, the heuristic approaches may seem more favorable than Brute-force in order to find an optimal subset of features from the sparse matrix, by assuming that the sheer number of dimensions will increase over time as new data arrive. Heuristics progressively improve a solution through iteration via a searching process. One of these searching processes is called Swarm search that looks via some probabilistic movements for a better solution at incremental steps. It then improves the overall solution towards the optimum, and is advantageous in solving NP-hard optimization problem. Instead of taking into consideration an enormously large set of combinations entirely, heuristics search and improve the solution at incremental steps.

Based on our proposed design, a new feature evaluation technique called clustering-by-coefficient-of-variation (CCV) can optimize the swarm search in OS-FS effectively. The replacement of the random start-up process by pre-allocated starting positions which are to be computed by CCV can optimize the initialization step of swarm search. Now, CCV is one of the favorable approaches in finding appropriate features for classification. Note that the features found by CCV are shown to be most appropriate for OS-FS to launch the swarm search.

To summarize, main advantages of the OS-FS method are:

- It may reach a close-to-optimum solution within a reasonable time frame;
- OS-FS has potentials for parallel processing.
- It outperforms the standard Best-Search in FS;
- OS-FS achieves higher accuracy without costing much extra time in most cases in comparison to other Swarm FS;

This paper is structured based on the following sections: "Literature review" provides a brief review on some recently developed feature selection methods which are claimed to be efficient as related work. In "Framework of OS-FS" section, the newly proposed feature selection method OS-FS is described. In "Experimental results", we present a computer simulation to illustrate our algorithm, with the results subsequently discussed in "Discussion". The conclusion is presented by "Conclusion".

Literature review

Tackling feature selection problem is a common issue in computer science. In practice, a more accurate classifier is observed when an appropriate feature subset is chosen. Previously, various researchers from the computer science communities pertaining to data mining, statistics and data science have attempted to find the best solution for this problem. It aims at yielding an optimal subset of features that are just sufficiently effective in enabling a maximum predictive power for the classifier. Searching for the optimal feature subset using the brute-force approach is computationally expensive due to the near infinite combinations. As an example, having 100 features can result in $2^{100} \approx 1.2677 \times 10^{30}$ subset combinations. In modern days, big database that contains millions of features are common [1].

In literature, streamlined feature selection methods include statistics-based methods dependent upon correlation, heuristics and metaheuristics as stochastic search from huge

search space of combinational subsets have been attempted [2]. The methods can give satisfactory results, and some metaheuristic search methods can break through the barrier of prohibitively large search space, but all the methods are computationally expensive. In addition, a longer time consumption versus a performance trade-off needs to be balanced [3].

Markov Blanket (MB) was designed to select features from large datasets to be tried as an efficient alternative of feature selection methods [4]. To ensure that all variables are probabilistically independent of the output feature subset, MB assumes a set of input variables. A subset of features can be derived from a Bayesian Network (BN) classifier based on MB of the class node. In accordance to graph theory, Markov blanket of a node n is the union of n's parents, n's children and the parents of n's children. This subset of nodes shields n from being affected by any node outside the blanket. Here, the MB of the class node denotes the selected features, and all other features outside the MB are deleted from the BN when extending the BN classifier on a complete and possibly larger dataset. Although the method's usage limits to only classifiers of Bayesian Network type, it is computationally efficient.

The clustering-based [5] and correlation-based [6] methods are the other types of fast feature selection methods that share a common advantage where the potentially huge search space for finding an ideal feature subset does not need to be fully searched. In order to pick the features that are strongly associating with the target classes, we examine only the values of the attributes and those of the target classes in pairwise or in clusters.

The fast clustering-based feature selection algorithm (FAST) [5] is the most recent advancement by the clustering-based technique that operates in two phases. By using graph-theoretic clustering methods, the features are grouped into clusters, after which the features are evaluated for their significance by relating them to the target classes; those strongly related ones are chosen from each cluster into the subset. For achieving fast processing speed a minimum-spanning tree clustering method is adopted. It is well known that the FAST is a fast and efficient technique similar to the correlation-based clustering algorithm [6] as it is based on the fundamentals of "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other" [7]. It assumes that the features that are strongly correlated to the target classes are indeed meaningful features even though the correlation principle may work well in most cases. The feature selections assume the features are independent when it comes to associating with the target classes. Some exceptions occur where the correlation principle may not succeed, and this also includes products of the feature selection methods. In fact, correlated features may pertain to noises, random or constant variables which correlate with the target classes occasionally. A feature that has little correlation with the target classes may be an important pairing factor to a key feature and should not be eliminated. A feature, namely *age* may seem to be a redundant variable, whereas pairing *age* and *gender* may give a different outcome. E.g., for $age > 15$ and female, this implies that the person is a potential cosmetics customer in classification model for marketing.

Based on our understanding, we examine the statistics of the dispersions—from there we separate between the qualified features and those otherwise using simple clustering method instead of relying on correlations between the features and the respective predicted classes. Swarm search is then applied to find an appropriate feature subset.

Framework of OS-FS

The OS-FS framework comprises of the following: The first part computes the scores of the features of the sparse matrix in accordance to principle of standard measure of dispersion. Qualified features are to be chosen as starting positions for swarm searching, which is supported by CCV. The second part is the swarm searching using selected candidates of starting positions derived from step 1 for activating swarm feature selection using wrapper feature selection with metaheuristics. The flowchart is illustrated by Fig. 2.

For finding suitable feature subset, a SS-FS framework depicts a wrapper type of feature selection and metaheuristic optimization [8]. The fitness function in the optimization part of the workflow that is on the left side of the execution flow is coded as the accuracy evaluator for a candidate classifier, whereby it tries to build using a candidate feature subset. As the optimization function attempts to search for a better subset in each round, iteration of its operation takes place. At every iteration, accuracy of the candidate classifier improves whenever a better subset can be found. The wrapped classifier serves as a fitness evaluator, and feedback the appropriateness of the candidate subset of features. Swarm search can be implemented by various metaheuristic search methods, which differ in their movement patterns, the way to converge and possibly how the search agents avoid falling into local optima. In SS-FS, our search agent is coded as a solution state comprising of a candidate subset of feature indices, and we assume variable size of feature subset. During initialization, we randomize the feature subset length. This is also applied to the choices of the features selected in the candidate subset. For $1 \leq k \leq K$ where K is the maximum cardinality of the search space that represents all the possible combinations of feature subsets given the largest dimension of the sparse matrix, we have the search agent store a variable k in the run-time memory. For particle swarm optimization

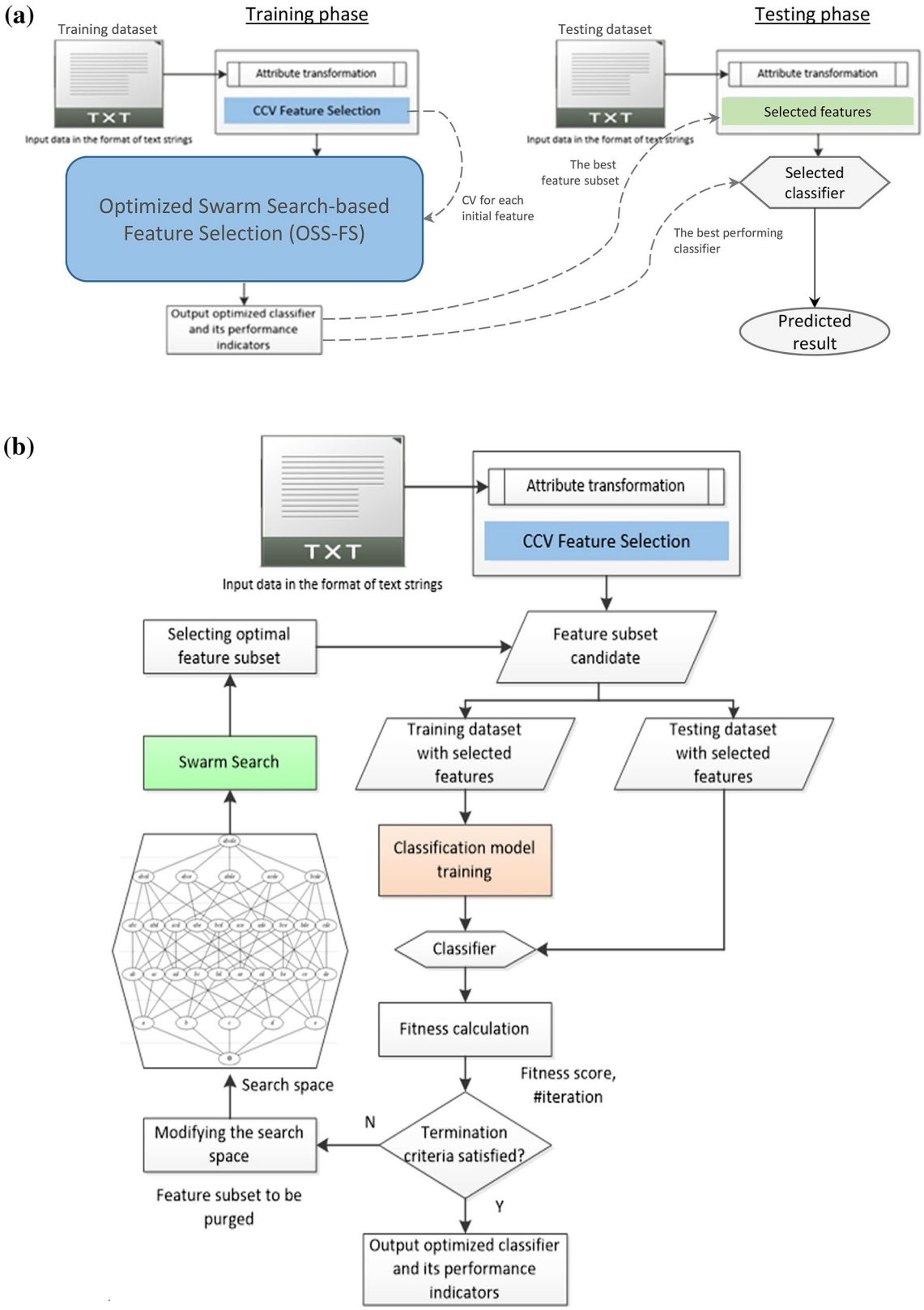

Fig. 2 **a** Operational flow of sentiment prediction by training then testing stages, **b** OSS-FS operation workflow

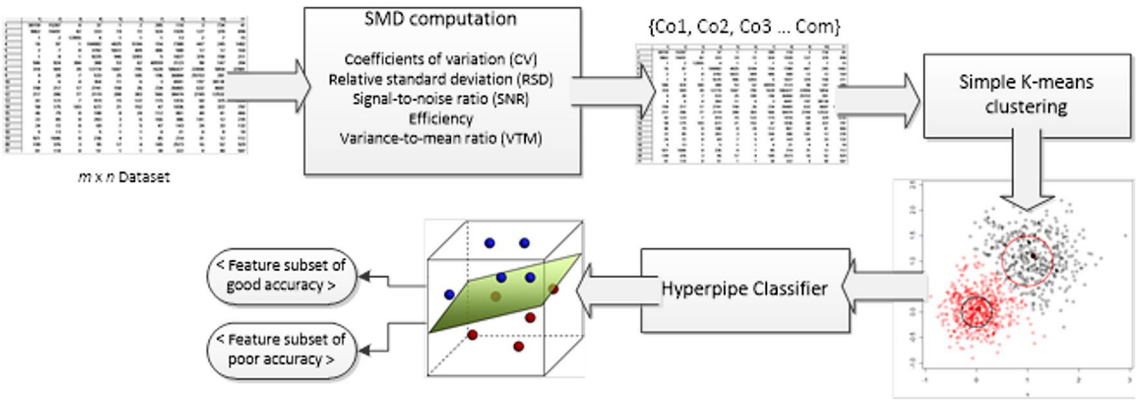

Fig. 3 Procedural steps of SMD feature selection framework

(PSO), the search agents move as a swarm across various dimensions such that search agents explore by their local velocities for doing local searches within their proximities, and globally they merge as a swarm towards the global optimum. The local searches are analogous to modifying the candidate feature subset by replacing some of the features with new candidates from the same dimension. When drawn by the global velocity that happen occasionally, the k values of the search agents modifies, for exploring new features from other dimensions than their current positions. Now, two modes of initialization are possible based on our framework. The starting positions of the search agents are randomly selected. The other mode is the main characteristic of OS-FS such that using the clustering-by-coefficients-of-variations (CCV) instead of random generation, the starting positions of the search agents are derived from the feature selection process over the full set of features sparse matrix [9].

Finding CCV seed positions for swarm search

Note that as an alternative feature selection approach, we implement CCV for fast and accurate running based on standardized measure of dispersion (SMD). By examining over the extent of dispersion for each feature, we cluster the coefficient values of dispersion into binary groups of useful and useless features. SMD includes coefficients of variation (CV), relative standard deviation (RSD), signal-to-noise ratio (SNR), efficiency and variance-to-mean ratio (VTM), and consists of a family of statistical methods for quantifying the dispersion into coefficient values. We are aware that CV is the fastest and implement it into our model. The operation flow of SMD comprises three steps (Fig. 3).

Firstly, we compute the coefficient values from the features by using one of the SMD statistics methods. The coefficient values as a single array are partitioned into two groups by simple K-means by their similarities due to its efficiency. Next, a group of features that can produce higher classification accuracy is selected.

We also know that the feature is significant enough to characterize a useful prediction model and the fact that a good attribute in a training dataset should have its data value vary sufficiently wide across a range of values. By using the Pima Indian diabetes dataset plotted by Projection Plot in Weka (Data Mining with Open Source Machine Learning Software), this occurence has been visualized [10]. It was observed from the visualized data pattern that the features that have a good distribution over the data space are those significant features in the classification model. Those attributes which do not vary much in the data scale and spread far in the data space will not be preferred.

Let us examine SMD in greater details. Regarding the extent of dispersion relative to the size of the observation, SMD has the advantage that the coefficients of dispersion are independent of the units of observation. By assigning X as the training dataset with n instances of vector whose values are characterized by a total of m attributes or features, we have an instance, m-dimensional tuple, in the form of $(x_1, x_2 \ldots a_m)$. For each x_a such that we have $a \in [1 \ldots m]$, one can then partition subgroups of different classes where $c \in C$ is the total number of prediction target classes so that $x_a \in \{x_a^1, x_a^2 \ldots x_a^c\}$. Now, we have in Eq. (1) such that

$$\text{CV}: \quad v_a = \sum_{c=1}^{C} \frac{\sqrt{\left[\sum_{j=1}^{n} \left(x_j^c - \overline{x_a^c}\right)^2\right]/n}}{\overline{x_a^c}} \tag{1}$$

where $\overline{x_a^c}$ is the mean of all the a^{th} feature values belonging to class c. Also, we have v_a as the sum of all coefficients of variation for each class c where $c \in [1 \ldots C]$ for that particular a^{th} feature. Coefficient of variation is a real number from $-\infty$ to $+\infty$.

The other variants: RSD, SNR, Efficiency, and VTM, are defined by the equations respectively.

$$\text{RSD}: \quad v_a = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{2}$$

$$\text{SNR}: \quad v_a = \frac{\mu}{\sigma} \tag{3}$$

$$\text{Efficiency}: \quad v_a = \frac{\sigma^2}{\mu^2} \tag{4}$$

$$\text{VTM}: \quad v_a = \frac{\sigma^2}{\mu} \tag{5}$$

where μ is the expected value, $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ is the sample mean, and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ is the sample variance, of feature data respective to the class c. These ratios are also known as standardized moments in statistics. They mainly measure how much a value deviates from its mean.

Our subsequent step is to find the threshold to decide what types of features and how many of them are to be retained after computing the coefficients values for each feature based on the Bias–Variance dilemma [11]. Some recent studies stated that the decomposition of error pertaining to a supervised learner into bias and variance terms allows substantial insight into the prediction performance of the classifier learner, which has its origins from analysis of regression which is the most basic form of classifier, the learning of models with numeric outputs. Squared *bias* measures the error of the central tendency of the classifier learner, whereas the *Variance* is a measure of the degree to which the classifier learner's predictions differ as it is applied to learn models from different training sets.

Note that the degree to which the predictions of those classifiers differ provides a lower limit on the average error of those classifiers when applied to subsequent test data if a learning system learns different concepts for different classifiers from different training sets. Now, inhibiting such variations between the classifiers will not necessarily eliminate prediction error despite the fact that the predictions from different classifiers differ. The degree to which the correct answer for an object can differ from that for other objects with identical descriptions ("irreducible error") and the accuracy of the learning bias also affects prediction error. Errors will also be caused by predictions from different classifiers that are the same but are not correct.

Next, the holdout approach of Kohavi and Wolpert [12] is a well-known employed approach for estimating bias and variance. Assume function: $t(x) = g(x) + \varepsilon$, the expected squared error over fixed size training sets D drawn from $P(X, T)$ can be expressed as the sum of three components:

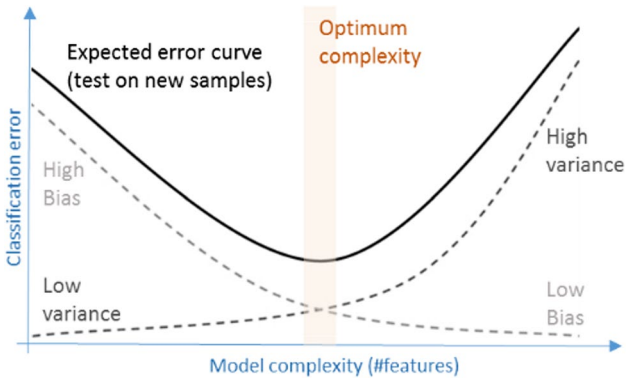

Fig. 4 Compromising of the error versus model complexity to achieve optimum complexity

$$\sum_{D}\left[\int_x\int_t (h(x)-t)^2 p(t|x)p(x)dtdx\right] = \sigma^2 + bias^2 + variance \tag{6}$$

$$\sigma^2 = unavoidable\ Error \tag{7}$$

$$bias^2 = \int\left(\sum_{D}[h(x)] - g(x)\right)^2 p(x)dx \tag{8}$$

$$\bar{h}(x) = \sum_{D}[h(x)] \tag{9}$$

$$variance = \int\sum_{D}\left[\left(h(x) - \bar{h}(x)\right)^2\right]p(x)dx \tag{10}$$

We have decomposed into the sum of a (squared) bias, a variance, and a constant noise term since our objective is to minimize the expected loss. Note that there is a trade-off between bias and variance, with very flexible models, which can over-fit, having low bias and high variance, and relatively rigid models (under-fit) having high bias and low variance. The URL: http://scott.fortmann-roe.com/docs/BiasVariance.html states as follows: "Managing bias and variance is really about managing over- and under-fitting. Bias is decreased and variance is increased in relation to model complexity. As more and more parameters are added to a classification model as descriptive features, the complexity of the model rises and variance becomes our primary concern while bias steadily falls."

As a demonstration, we present Fig. 4, which illustrates the two contradicting trends on the increases/decreases of Bias and Variances.

If the total error is at minimum, and the variance and bias curves intersect, and our model complexity exceeds this ideal spot that exist as an optimum area near the intersection between these three curves, we are in effect over-fitting our

Table 1 Algorithms implementation and experimental setup

Algorithm	Method type	Refs.	Description of method
SMO	Classifier	[14]	Using scaled polynomial kernels, a support vector machine is induced by this sequential minimal optimization algorithm. The SVM outputs are transformed into probabilities via a sigmoid function, for deciding the membership of the target class when the output does not fit the data
SGM	Classifier	[15]	Sparse Generative Modelling, it is designed for scalable and accurate text classification. It is claimed to achieve reduced time complexity because of fast inference using a sparse model representation combined with the use of inverted index
Cfs-SubsetEval	Feature evaluator	[16]	By ensuring the correlation strength between a candidate subset of features and the target class is high, and the inter-correlations with the other classes are low, it estimates the worth of a subset of attributes. In this method, correlation with the class is taken as the predictive ability for each feature in the subset
Consistency-SubsetEval	Feature evaluator	[17]	By checking the extent of consistency in the class values while a candidate feature subset is projecting on the training instances, it evaluates the worth of the feature subset
Best-search	Search method	[18]	It implements the best-first search strategy based on hill climbing to navigate feature subsets
EvolutionarySearch-random-selection	Search method	[19]	It explores the feature search space using an Evolutionary Algorithm. Population size = 20, Mutation probability = 0.01, and Selection Operator = Random
EvolutionarySearch-tournament-selection	Search method	[20]	It explores the feature search space using an Evolutionary Algorithm. Population size = 20, Mutation probability = 0.01, and Selection Operator = Tournament
PSOSearch-bit-flip	Search method	[21]	It explores the feature search space using Particle Swarm Optimization Algorithm. Population size = 20, Individual weight = 0.34, Inertia weight = 0.33, Mutation probability = 0.01, and Mutation Type = Bit-flip

model. On the other hand, we are under-fitting the model if our complexity falls short of the necessary features. In fact, there is no easy analytical way to find the ideal location.

We have a simple clustering technique in order to achieve this optimum equilibrium and attempt to partition the statistical scores of the features into a small number of clusters, which may be consistent in terms of similarity among its members. Typically, S scores of the statistical coefficients s_i, with indices $i = 1 \ldots S$ have to be partitioned into two clusters. Our goal is to assign membership of a cluster to each score, and we have to find the ideal cluster positions μ_i, $i = 1 \ldots k$ of the clusters that minimize the distance from the data points to the cluster centroids. Our objective function is as follows:

$$\begin{aligned} arg &= \min_{p} \sum_{i=1}^{2} \sum_{s \in p_i} d\left(s, \mu_i\right) \\ &= arg \min_{p} \sum_{i=1}^{2} \sum_{s \in p_i} s - {\mu_i}^2 \end{aligned} \tag{11}$$

where p_{i} is the partition of scores that belong to cluster i. The clustering algorithm implements square of Euclidean distance as: $d\left(s, \mu_i\right) = s - {\mu_i}^2$

A simple method is used here to find a spot that is kept very close to the ideal location. We define the variance as a measure of the contribution to error of deviations from the central tendency under which every instance is regarded as a spot in multi-dimensional space, and divide it into attribute parts. Next, we can calculate the score of dispersion of every attribute by computing the variance of every attribute, coefficient of variation (or other statistical definitions of dispersion under SMD) and variance that has multiple relationships. Given a data set $X = \{x_1, x_2, \ldots, x_n\}$, the estimation function is: $f(x) = a_o x_{0+} a_1 x_{1+} a_2 x_{2+} \ldots = \vec{a}\vec{x}$. As we have known, adding more parameters into the model as features, the complexity of the model rises, so does the variance while bias drops. The function of K-means is to divide the data set into two groups according to the values of coefficient of variation. The values of variance-bias are different for the data points in different clusters that reflect the complexity of model. It is known the more a complex model, the more bias it is, and vice versa. Therefore, reducing the complexity of model by choosing some valuable attributes by separating the variance is achieved. Note that total errors of the groups are given by the following equations:

$$cluster_1 = bias^2 \uparrow + variance \downarrow + \sigma^2 \tag{12}$$

Table 2 Sentiments of the meaningful use functionality items [22]

Meaningful use functionality	Number of MU impacts	Positive (%)	Mixed-positive (%)	Neutral (%)	Negative (%)
Clinical decisions support	142	65	17	11	7
Computerized provider order entry	91	63	16	12	9
Multifunctional health IT intervention	131	51	33	8	8
Health information exchange	33	64	30	0	6
e-Prescribing	25	52	28	4	16
Patient lists by condition	30	73	17	3	7
Patient access to electronic records	20	60	25	10	5
Patient care reminders	10	60	30	0	10
Other meaningful use functionalities	11	55	36	9	0

$$cluster_2 = bias^2 \downarrow + variance \uparrow + \sigma^2 \quad (13)$$

One of the two clusters by Eq. (12) and Eq. (13) with scores of dispersions representing the combinations of variances and biases is to be chosen as the optimal feature subset. Hyper-Pipes [13] is utilized for this task, which is a probabilistic learning tool that is very similar to Naïve Bayes except that it does not record the frequency count of how attributes correspond to classes. In fact, an attribute either corresponds to a hypothetical class or it does not, regardless of how frequent it occurs. We then record all of the attributes and their correspondence with the class in a table of Booleans, and determine the class based on the score of the attributes added up (0 for existence, 1 for non-existence).

Experimental results

We have two popular feature evaluators, four search algorithms and three optimized search algorithms for verifying the efficacy of OS-FS and we conduct the experiment using two good performing classifiers. The three optimized search algorithms, under the framework of OS-FS are improved versions of the three out of four search algorithms, taking Best-Search as a comparison baseline. The optimized search algorithm coefficient-of-variation was applied in scoring the features. The algorithms are listed (Table 1).

The training data are excerpted from a review survey report called "Effects of Meaningful Use Functionalities on Health Care Quality, Safety, and Efficiency"[3] which is released in 2014. The report reviews the January 2010 to August 2013 health IT literature to examine the effects of health IT across three aspects of care: efficiency, quality, and safety, and updates previous systematic reviews of the health IT literature. In particular, it focuses on identification and summary of the data related to the use of health IT, which was outlined in the Meaningful Use regulations. This review examined previous literature in order to derive article authors' findings that have relation to the effects or associations of a meaningful use functionality on an aspect of care. Each article's findings was scored as (1) positive, which is defined as health IT improved key aspect of care but none worse off; (2) mixed-positive, which is defined as positive effects of health IT outweight negative effects; (3) neutral, which is defined as health IT not associated with change in outcome; and (4) negative, which is defined as negative effects of health IT on outcome.

Based on the input from a panel of five nationally-known health IT experts, this systematic review was performed using three stages by health IT subject matter experts, who utilize web-based system to conduct screening. The first stage involved independent, dual-rater screening of articles based on their titles against a set of defined on the inclusion/exclusion criteria. The second stage involved screening each article at the abstract level using a standardized abstraction form. The third stage of the screening process involved a full text review and classification using a standardized abstraction form. Composition of the 'Meaningful Use Functionality' and the corresponding sentiment assessed by the reviewers can be shown by Table 2.

The training dataset is first formatted from raw text into ARFF format (a proprietary file format for Weka). The formatted file consists of one news per entry, whereas each entry has a dual structure of {*text-string, emotion-label*} of varying string length. Figure 3 shows the training dataset that is subjected to the OS-FS process.

Referring to Fig. 3, the relations between classifier, feature evaluator and search method are:

1. Classifier serves as a fitness function, input by a candidate subset of features and output some performance indicator, e.g. Accuracy. The classifier that is induced

[3] http://hitconsultant.net/2014/03/05/onc-releases-report-effects-meaningful-use-functionalities-healthcare-quality-safety-efficiency/.

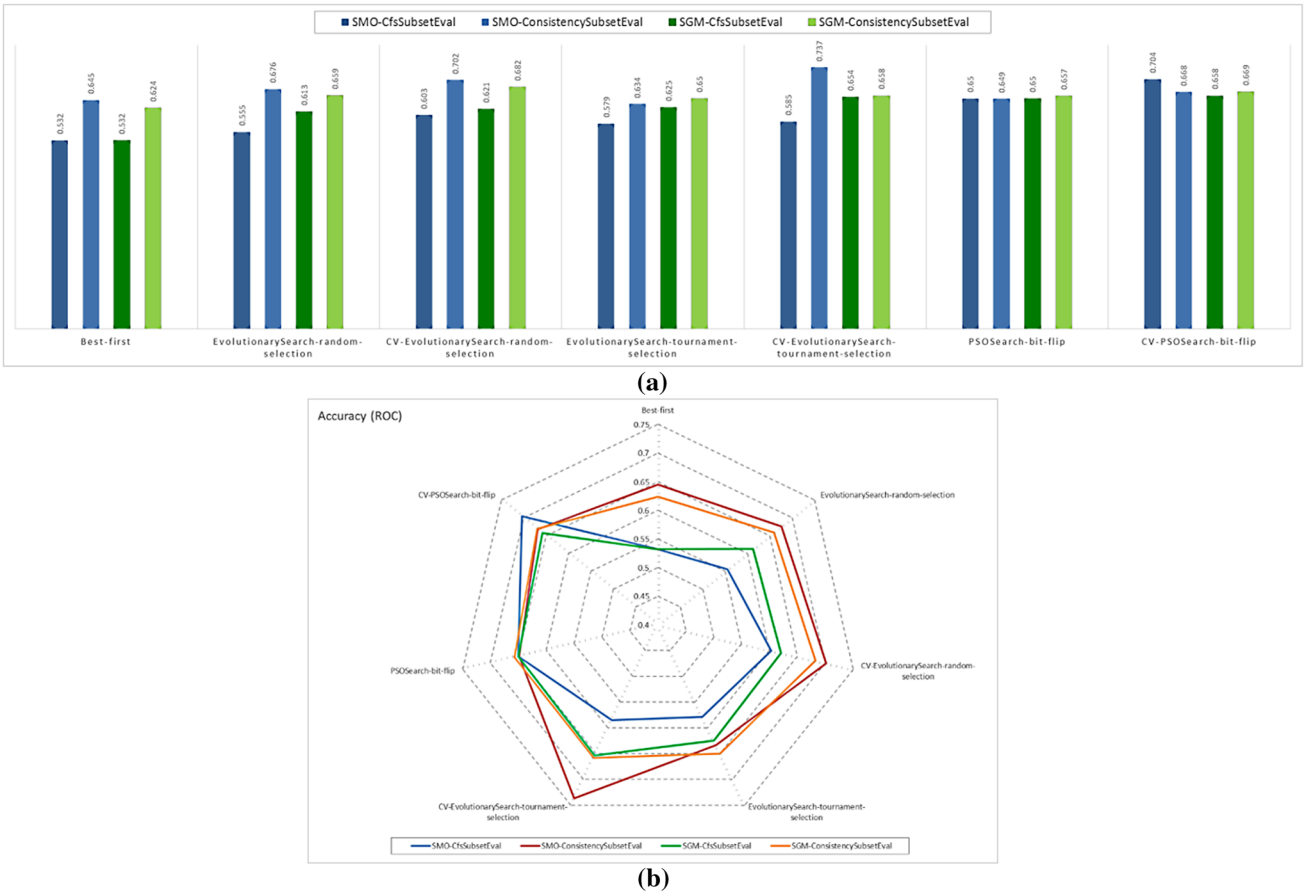

Fig. 5 Bar-chart (**a**) and radar-chart (**b**) depicting accuracy based on different combinations of classifiers, feature evaluators and search methods

in the final round of stochastic optimization would be the deliverable of the OS-FS as it should be built by the most suitable subset after rounds of searches.

2. Feature evaluator takes input from search method in the form of candidate feature subset; it performs a preliminary check over the candidate feature subset, validating on whether the candidate subset is qualified by the standard of the feature evaluator (e.g. correlation or consistency). If it qualifies, the candidate feature subset is passed onto inducing a new version of classifier. Search method depicts how the search agents move around the search space looking for a better subset.

The data is processed with the aim of comparing the feature selection methods with and without the use of CCV. The feature selection methods enhanced by CCV carry a prefix "CV-". Two types of classifiers, which are the SMO and SGM, are induced by using combinations of feature selection evaluators and search methods. The performance for evaluation are: (1) accuracy in the form of ROC, and (2) Kappa statistics. Note that ROC is the abbreviation for receiver operating characteristic curve that is a plot of the true positive rate against the false positive rate. The accuracy defined here is the area under the curve as a measure of text accuracy, which a maximum value 1 (highest predictive power) and minimum zero, which can be taken as random.

The configuration of the experiment platform is the Intel Core, i7-4785T CPU @ 2.2 GHz, 8 GB RAM, Windows 8.1 and 64-bit Operating System, x64-based processor, and the data mining software is Weka 3.7. The classification model is built by using 10-fold cross validation, which can ensure a stringent performance evaluation with 10 different proportions of train/test instances cross-validating one another.

For different methods listed in Table 1, the combinations of experiment runs are combined from two main groups of classifiers—one is by SMO and another by SGM. For each group of classifier, two feature evaluators, which are given by Cfs and consistent, are paired with a total of seven search methods. The performance results that run under Weka are charted in Figs. 5 and 6 in terms of Accuracy (ROC) and Kappa Statistics respectively.

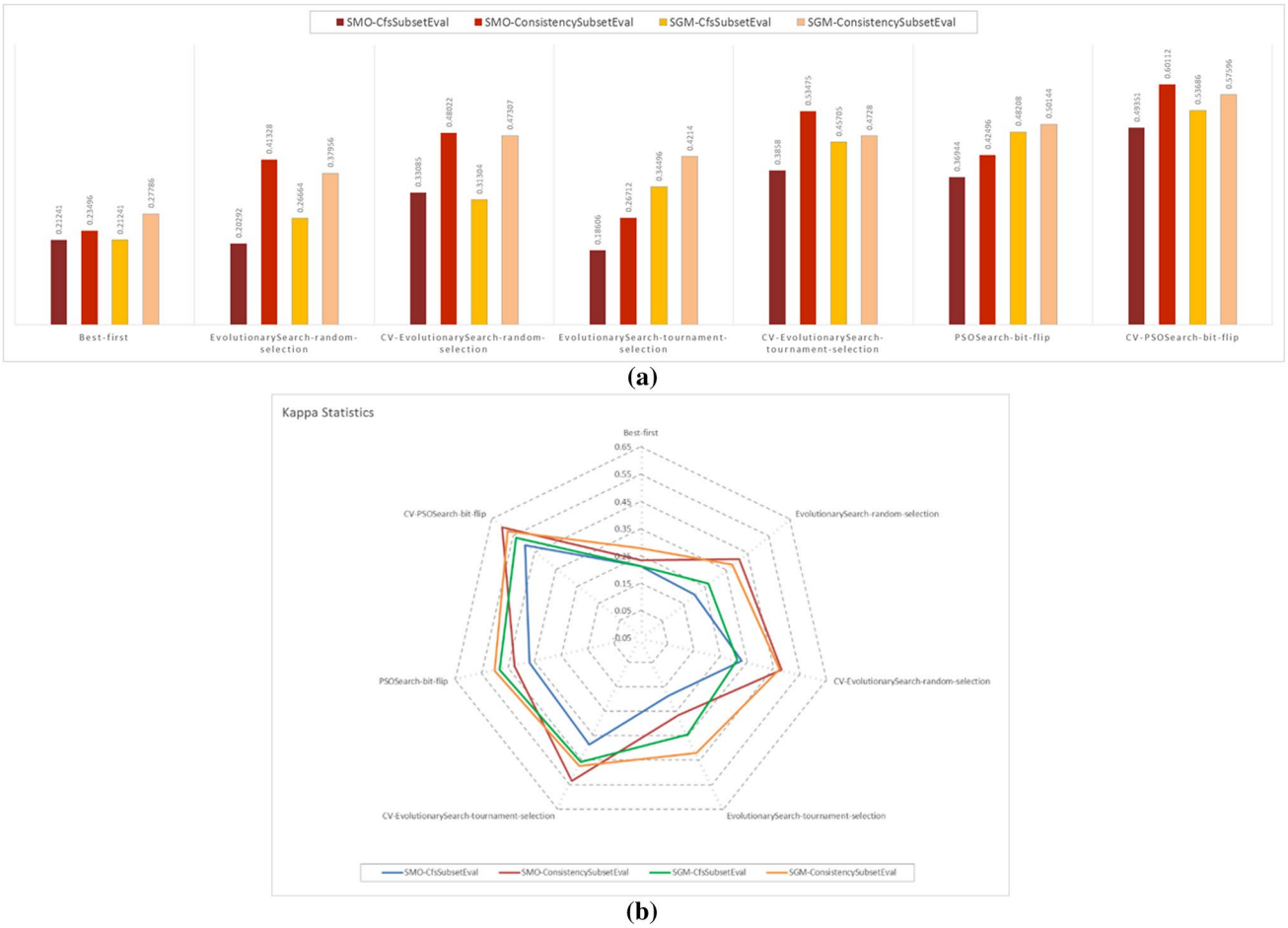

Fig. 6 Bar-chart (**a**) and radar-chart (**b**) depicting Kappa Statistics based on different combinations of classifiers, feature evaluators and search methods

Discussion

Note that two phenomena can be consistently seen in all cases of the experimentation. First of all, the swarm-based feature selection methods outperformed Best-first when coupled with SGM, which is the baseline and is supposed to be the best of all the non-swarm search methods in Weka, which indicates that swarm search methods do have their certain merits over feature selection. This is because they are able to heuristically optimize the subset along with iteration. There is a tie as swarm-based FS made 3-wins and 3-loses comparing to best search in the case of SMO, which shows that SGM is a more appropriate method as the model is inferred from simple and scalable probabilistic Bayesian matrix. We also point out that SMO is based on SVM with an optimized kernel. The results by SMO tends to be unstable when the kernel alignment happens in high dimensions of state search space.

The second observation from the bar-charts is that the CV-based versions of swarm search outperformed all those without CV enhancement, which is an important observation in this paper since it demonstrates that the concept of starting the search positions by the information of CV does have an edge in the feature selection performance. This effect is more obvious in cases of SGM than in cases of SMO, which implies that SGM is an appropriate choice of classification algorithm to incorporate with OS-FS. For the SGM versus the SMO, accuracy gains for CV based methods are generally higher. For example, CV-EvolutionarySearch-Tournament-selection achieves up to 73.7% in accuracy, which is the highest in this experimentation, for ConsistencySubsetEval with either SGM or SMO classifier. The 73.7% accuracy may be relatively low, but considering the challenges in text mining where very high accuracy is hard to attain.

The feature evaluator, Consistency tends to be better that the Cfs. Consistency based FS evaluator is able to achieve higher accuracy while retaining more features than Cfs that was designed to minimize the length of feature subset. In addition, the search times by Consistency based FS evaluator tends to be much shorter than those by Cfs.

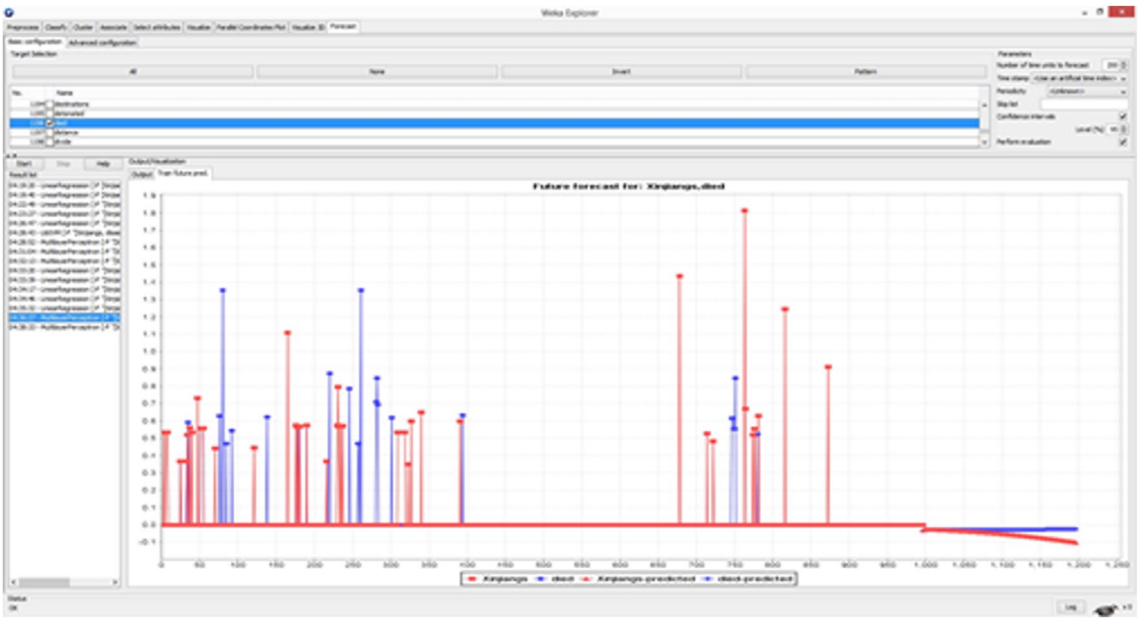

Fig. 7 Time series forecasting of future trends pertaining to medical articles based on the cleansed dataset after application of dimensionality reduction

author	article_title	publication_year	meaningful_use_functionality	aspect_of_care	author_sentiment
Abdel-Qader DH, Harper L, Cantrill JA, et al	Pharmacists' interventions	2010	e-Prescribing	safety	negative
Ali J, Barrow L, Vuylsteke A	The impact of computerised physician	2010	Computerized Provider Order Entry	safety	positive
Bell LM, Grundmeier R, Localio R, et al	Electronic health record-based decisio	2010	Clinical Decision Support	quality	positive
Bennett KJ, Steen C	Electronic medical record customizatic	2010	Other Meaningful Use	efficiency	positive
Bernstein JA, Imler DL, Sharek P, et al	Improved physician work flow after in	2010	Multifaceted health IT Intervention	quality	positive
Bernstein JA, Imler DL, Sharek P, et al	Improved physician work flow after in	2010	Multifaceted health IT Intervention	efficiency	positive
Bourgeois FC, Linder J, Johnson SA, et al	Impact of a computerized template on	2010	Clinical Decision Support	quality	mixed-positive
Co JP, Johnson SA, Poon EG, et al	Electronic health record decision supp	2010	Clinical Decision Support	quality	positive
Damberg CL, Shortell SM, Raube K, et al	Relationship between quality improve	2010	Multifaceted health IT Intervention	quality	neutral
Davis AM, Cannon M, Ables AZ, et al	Using the electronic medical record to	2010	Patient Lists By Condition	quality	positive
Dejesus RS, Angstman KB, Kesman R, et al	Use of a clinical decision support syste	2010	Clinical Decision Support	quality	positive
DesRoches CM, Campbell EG, Vogeli C, et al	Electronic health records' I	2010	Multifaceted health IT Intervention	quality	mixed-positive
Devine EB, Hansen RN, Wilson-Norton JL, et al	The impact of computerized provider c	2010	Computerized Provider Order Entry	safety	positive
Devine EB, Hollingworth W, Hansen RN, et al	Electronic prescribing at the point of c	2010	e-Prescribing	efficiency	negative
Downs SM, Anand V, Dugan TM, et al	You can lead a horse to water: physicia	2010	Clinical Decision Support	quality	negative
Duffy RL, Yiu SS, Molokhia E, et al	Effects of electronic prescribing on the	2010	e-Prescribing	quality	positive
Duffy RL, Yiu SS, Molokhia E, et al	Effects of electronic prescribing on the	2010	e-Prescribing	efficiency	mixed-positive
Duffy WJ, Kharasch MS, Du H	Point of care documentation impact or	2010	Multifaceted health IT Intervention	efficiency	mixed-positive
Feldstein AC, Perrin NA, Unitan R, et al	Effect of a patient panel-support tool c	2010	Patient Lists By Condition	quality	positive
Fiumara K, Piovella C, Hurwitz S, et al	Multi-screen electronic alerts to augm	2010	Clinical Decision Support	quality	positive
Furukawa MF, Raghu TS, Shao BB	Electronic medical records, nurse staff	2010	Multifaceted health IT Intervention	quality	mixed-positive
Furukawa MF, Raghu TS, Shao BB	Electronic medical records, nurse staff	2010	Multifaceted health IT Intervention	efficiency	negative
Furukawa MF, Raghu TS, Shao BB	Electronic medical records, nurse staff	2010	Computerized Provider Order Entry	efficiency	negative
Galanter WL, Thambi M, Rosencranz H, et al	Effects of clinical decision support on v	2010	Clinical Decision Support	quality	positive
Guerra YS, Das K, Antonopoulos P, et al	Computerized physician order entry- b	2010	Computerized Provider Order Entry	quality	positive
Hill PM, Mareiniss D, Murphy P, et al	Significant reduction of laboratory spe	2010	Computerized Provider Order Entry	safety	positive
Himmelstein DU, Wright A, Woolhandler S	Hospital computing and the costs and c	2010	Multifaceted health IT Intervention	quality	mixed-positive
Himmelstein DU, Wright A, Woolhandler S	Hospital computing and the costs and c	2010	Multifaceted health IT Intervention	efficiency	neutral
Hoekstra M, Vogelzang M, Drost JT, et al	Implementation and evaluation of a n	2010	Clinical Decision Support	quality	positive
Holt TA, Thorogood M, Griffiths F, et al	Automated electronic reminders to fa	2010	Patient Lists By Condition	quality	neutral
Holt TA, Thorogood M, Griffiths F, et al	Automated electronic reminders to fa	2010	Patient Lists By Condition	quality	mixed-positive
Jones SS, Adams JL, Schneider EC, et al	Electronic health record adoption and	2010	Multifaceted health IT Intervention	quality	mixed-positive
Kesman RL, Rahman AS, Lin EY, et al	Population informatics-based system f	2010	Patient Lists By Condition	quality	positive
Ling SB, Richardson DB, Mettenbrink CJ, et al	Evaluating a Web-Based Test Results S	2010	Patient Access to Electronic Records	efficiency	positive
Longhurst CA, Parast L, Sandborg CI, et al	Decrease in hospital-wide mortality ra	2010	Computerized Provider Order Entry	quality	positive
Mattison ML, Afonso KA, Ngo LH, et al	Preventing potentially inappropriate r	2010	Computerized Provider Order Entry	safety	positive
Mayer PH, Yaron M, Lowenstein SR	Impact on length of stay after introduc	2010	Computerized Provider Order Entry	efficiency	neutral
McCluggage L, Lee K, Potter T, et al	Implementation and evaluation of van	2010	Computerized Provider Order Entry	quality	positive
McCoy AB, Waitman LR, Gadd CS, et al	A computerized provider order entry i	2010	Computerized Provider Order Entry	safety	positive
McCullough JS, Casey M, Moscovice I, et al	The effect of health information techn	2010	Multifaceted health IT Intervention	quality	mixed-positive
Metzger J, Welebob E, Bates DW, et al	Mixed results in the safety performan	2010	Computerized Provider Order Entry	safety	negative
Moore LJ, Turner KL, Todd SR, et al	Computerized clinical decision suppor	2010	Clinical Decision Support	quality	positive

Fig. 8 A snapshot of sample training dataset pertaining to medical articles

The other performance indicator Kappa that measures how generalized a classification model is when it comes to work on other datasets than the training dataset can be interpreted as the "reliability" [23] of a classifier. It has a decimal value and we aim to have it as high as possible. Figure 6 illustrates that the Kappa statistics are all improved when CV are used in the swarm search methods. This implies that classifiers that are built using CV methods are more robust when they come to be tested with other unseen testing data, and the corresponding Kappa values for SGM are better than those for SMO.

We have an example illustrated by Figs. 7 and 8 as an extended scope of this research. The figure predicts the future trends of news articles carrying the keywords 'e-Prescribing' and 'Clinical Decision Support' respectively. An artificial neural network is used as the base predictor. The trends are showing a forecasted slight decline of news involved the keyword 'outbreak', as well as a deeper decline of news involving 'epidemic' pertaining to negative emotion.

Conclusion

For finding the right feature subset in a huge state-space search when the data are highly dimensional, the feature selection technique has long been considered as the ideal dimensionality reduction problem in text mining. It is well known that when text strings are transformed into sparse matrix, a very large set of features are expected after the string-to-vector transformation. The search space is discrete but it contains all possible combinations of features you could select from the dataset. Swarm search methods have been recently proposed as an effective mean to navigate through the search space and discover a close to be best combination of features that improves classifier performance over using all features.

We proposed the OS-FS concept and supported it with experiments based on the principles of balancing between overfitting and underfitting by the CCV principle. The design of OS-FS is aimed at achieving fast FS, so the speed of swarm-based searches would not be comprised, as swarm search usually requires longer run-time because of its iterative (stochastic) characteristic. The features by CCV are first produced as a pre-processing step for swarm-based FS, suggesting the appropriate seeding positions for the search agents. The proposed scheme is validated via a sentiment classification experimentation given that 279 instances of medical articles are extracted from MEDLINE as training/testing samples. Authors' emotions (or sentiments) are used as target class labels. Our results show that OS-FS can improve the default swarm-based FS in Weka with certain gain in both accuracy and Kappa. Best-search that represents the base-line of non-swarm search method for FS is outperformed by swarm-type FS and enhanced versions of swarm-type FS algorithms. Our improvement in the algorithm can lead to better classifiers for medical text mining based on the sentiment prediction.

As future work, it is intended to apply the proposed dimensionality reduction techniques into building an automatic sentiment prediction system. When a suitable classifier has been sufficiently trained, it can be deployed to automatically scout and forecast the anticipated sentiment of the reader mass from a particular news. Being able to predict the collective emotion of the audience mass would be useful for sentiment prediction applications, such as election prediction, stock market prediction and product hype cycle forecast etc. On the other hand, our proposed dimensionality reduction method when coupled with filtering technique of mis-classified instances removal, it can potentially produce a 'clean' training dataset where the data and features of ambiguity and poor predictive power are cleansed. Hence the cleansed data can be used for time-series forecasting that offers foresights of a life-time of news with certain keywords.

Acknowledgements This paper is supported by the research grant "Temporal Data Stream Mining by Using Incrementally Optimized Very Fast Decision Forest (iOVFDF)," Grant No. MYRG2015-00128-FST, which is offered by the University of Macau, FST, and RDAO.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ehtical approval This article does not contain any studies with human participants and animals performed by any of the authors.

References

1. Lakshminarayan CK (2013) High dimensional big data and pattern analysis: a tutorial. In: Bhatnagar V, Srinivasa S (eds) Big data analytics, Lecture Notes in Computer Science, Springer, Cham. https://doi.org/10.1007/978-3-319-03689-2_5
2. Yusta SC (2009) Different metaheuristic strategies to solve the feature selection problem. Pattern Recognit Lett 30(5):525–534. https://doi.org/10.1016/j.patrec.2008.11.012
3. Fong S, Deb S, Yang XS, Li J (2014) Feature selection in life science classification: metaheuristic swarm search. IEEE IT Prof 16(4):24–29. https://doi.org/10.1109/MITP.2014.50
4. Tsamardinos I, Aliferis CF, Statnikov A (2003) Time and sample efficient discovery of markov blankets and direct causal relations. In Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, pp. 673–678
5. Song Q, Ni J, Wang G (2013) A fast clustering-based feature subset selection algorithm for high-dimensional data, IEEE

Trans Knowl Data Eng 25(1):1–14. https://doi.org/10.1109/TKDE.2011.181
6. Baris S (2008) Fast correlation based filter (FCBF) with a different search strategy. In Proceedings of 23rd international symposium on computer and information sciences, IEEE, Oct. 2008, pp. 1–4
7. Hall MA, Smith LA (1999) Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In Proceedings of the 12th international florida artificial intelligence research society conference, pp. 235–239
8. Fong S, Deb S, Yang X-S, Li J (2014) Metaheuristic swarm search for feature selection in life science classification. IEEE IT Prof 16(4):24–29
9. Fong S, Liang J, Wong R, Ghanavati M (2014) A novel feature selection by clustering coefficients of variations. In: 2014 ninth international conference on digital information management (ICDIM), 29 Sep–1 Oct 2014, pp. 205–213
10. Fong S, Liang J, Deb S (2013) Diabetics prediction by using feature selection based on coefficient of variation. In: Proceedings of Wilkes—international conference on computing sciences, New Delhi, November 2013
11. Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. Neural Comput 4(1):1–58
12. Hassanien A-E, Azar T, Snásel A, Kacprzyk V, Abawajy J, J.H. (eds) (2015) Big data in complex systems: challenges and opportunities. Studies in Big Data. Springer, Cham
13. Muskan Kukreja SA, Johnston, Stafford P (2012) Comparative study of classification algorithms for immunosignaturing data. BMC Bioinf 13:139
14. Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In: Scholkopf B, Burges C, Smola A (eds) Advances in kernel methods: support vector learning. MIT Press, Cambridge
15. Jacob Eisenstein A, Ahmed, Xing EP (2011) Sparse additive generative models of text. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 1041–1048
16. Hall MA (1998) Correlation-based feature subset selection for machine learning, PhD thesis, University of Waikato, Hamilton, New Zealand
17. Liu H, Setiono R (1996) A probabilistic approach to feature selection—a filter solution. In: 13th international conference on machine learning, pp. 319–327
18. Ohta K, Moriai S, Aoki K (1995) Improving the Search Algorithm for the Best Linear Expression. Advances in cryptology—CRYPT0′95, Lecture Notes in Computer Science, vol 963, pp. 157–170
19. Ferrer J, Kruse PM, Chicano F, Alba E (2015) Search based algorithms for test sequence generation in functional testing. Inf Softw Technol 58:419–432
20. Bravo Y, Luque G, Alba E (2015) Takeovers time in evolutionary dynamic optimization: from theory to practice. Appl Math Comput 250(1):94–104
21. Moraglio A, Di Chio C, Poli R (2007) Geometric Particle Swarm Optimisation. In: Proceedings of the 10th European Conference on Genetic Programming, Berlin, Heidelberg, pp. 125–136
22. Jones SS, Rudin RS, Perry T, Shekelle PG (2014) Health information technology: an updated systematic review with a focus on meaningful use. Ann Intern Med 160(1):48–54
23. Fong S, Zhang Y, Fiaidhi J, Mohammed O, Mohammed S (2013) Evaluation of stream mining classifiers for real-time clinical decision support system: a case study of blood glucose prediction in diabetes therapy. Biomed Res Int. https://doi.org/10.1155/2013/274193