JOURNAL OF PURE AND APPLIED MICROBIOLOGY, Aug. 2015. Vol. **9**, p. 139-146
*SPECIAL ISSUE ON RECENT RESEARCH CHALLENGES IN BIO-MEDICAL APPLICATIONS*

# Fish Swarm Optimization for Feature Subset Selection in Medical Information Blog Opinion Mining

**K.M. Subramanian[1]* and K. Venkatachalam[2]**

[1]Department of Computer Science Engineering,
Erode Sengunthar Engineering College, Erode, Tamilnadu, India.
[2]Department of Electronics and Communication Engineering,
Velalar College of Engineering and Technology, Erode, Tamilnadu, India.

(Received: 10 February 2015; accepted: 28 April 2015)

**With search for health related information increasing in the web and availability of various blogs for medical information, analysis of subjective information is challenging. Opinion Mining is an emerging area, which labels an opinion as positive or negative. Opinion mining has been extensively used in product / movie reviews and to the best of our knowledge has not been investigated on data collected from medical question and answer blogs. Feature selection for labelling is challenging, due to the various medical terminology used. For example, the common cold is also called rhino pharyngitis, upper respiratory tract infection or naso-pharyngitis. This work investigates a novel feature selection technique using Fish Swarm Optimization, which identifies key medical concepts found in blogs and labels whether the opinion of the patient is positive or negative for the treatment undertaken.Three classification algorithms Naïve Bayes, K Nearest Neighbour and Classification and Regression Trees (CART) algorithms were used to investigate the quality of the extracted features.**

**Key words:** Opinion Mining, Principal Component Analysis (PCA), Kernel PCA, Fish Swarm Optimization (FSO).

Opinion is what a person thinks about something. It detects and extracts subjective information in text documents[1]. Opinions are subjective expressions describing people's sentiments/appraisal/feelings to entities/events/properties[2].Opinion mining studies opinions at word level, sentence level and document level[3].It is a new discipline, which attracted attention in fields like marketing, personal affective profiling and financial market prediction.

Opinion mining (OM), also called Sentiment classification or Polarity classification, is a binary classification task labelling an opinionated document expressing either overall positive or overall negative opinion. A technique to analyze subjective information in many texts and studies is sentiment classification. An approach to sentiment classification is using machine learning algorithms. Sentiment analysis tasks include classifying a text polarity at document, sentence or feature/aspect level expressing opinions, which are positive, negative or neutral. Sentiment analysis is performed at document, sentence and feature levels[4].

Feature extraction identifies/selects sufficient features set to characterize a texture. Image coding provides a compact texture description from chosen features. By representing a complex texture with limited measurable features/parameters, texture analysis achieves dimension-reduction enabling automated texture processing[5].

Feature Selection (FS) identifies significant features and eliminates irrelevant/dispensable ones. FS, also called feature subset

* To whom all correspondence should be addressed.
E-mail:

selection, variable selection, or attribute reduction, is common in machine learning. Basically, it selects a dataset's input attributes that closely defines a specific outcome[6].Feature selection saves a learning procedure's operating time eliminating irrelevant/redundant features. Learning algorithms without intervention of irrelevant, redundant, and noisy features can centrally point out essential data features and build simpler and precise data models. Feature selection builds a simpler and common model and gets better insight into a task's fundamental perception[7].

This work investigates feature extraction methods and classification algorithms efficacy to classify medical opinions reviews. Two conditions insomnia and common cold related blogs were selected from various online health forums and the impact of specific type of treatment (specific drug, yoga, alternative medicine) was positive or negative in curing the condition were rated by a health professional. Feature transformation was achieved using Principal Component Analysis (PCA) and kernel PCA with the best features selected using Fish Swarm Optimization (FSO). Naïve Bayes, K Nearest Neighbour (KNN) and CART algorithms performance evaluation were evaluated for the proposed feature selection method. The paper is organized into the following sections. Section 2 reviews some related works in literature. Section 3 contains various techniques used in the investigation and Section 4 discusses the results obtained.

**Related works**

Counting positive and negative term scores to resolve sentiment orientation was undertaken by Ohana and Tierney[8] with improvement being presented by building a relevant feature dataset using SentiWordNet as source, and applying it to a machine learning classifier. Results through SentiWordNet corresponded to similar approaches using manual lexicons in literature. Feature sets approach also improved over baseline term counting procedures. Results from simple word counting were the same as other results using manual lexicons, showing that SentiWordNet performed well incomparison to manual task resources.

The role of BGDs among sentence words for OM and many BGDs that directly facilitate OM were explored by Srivastava *et al*.,[9]. Due to the unavailability of universal resources (opinion word lexicon/feature corpus) for application domains, OM was challenging. Free format product reviews OM mainly dealt with domain specific features extraction, opinion words identification corresponding to features and determination of Semantic Orientation of word opinion (large, small, loud, etc.), which do not belong to set prior polarity opinion words (beautiful, excellent, good,).

A newmethod to deal with feature-level OM problems was proposed by Liu *et al*.,[10]. Specially, 1) the new method considered explicit and implicit features. 2) Opinion words were split into 2 categories, vague opinion words and clear opinion words, for implicit features identification and clustering features. Feature clustering was based on three aspects: corresponding opinion words, similarity of features and features structures. Also, context information enhances clustering in the process as proved useful in clustering. Experiments demonstrated that the new method performs better.

A solution to sentiment classification, when target domain lacked labels, was developed by Pan *et al*.,[11]. However, there was some labeled data in different domains, regarded as source domain. In such cross-domain sentiment classification setting, to bridge inter-domain gap, a Spectral Feature Alignment (SFA) algorithm was tried to align domain-specific words from various domains to unified clusters aided by domain-independent words as bridge. SFA discovers a robust cross-domain data representation by exploiting relationship between domain-specific and domain-independent words through simultaneously co-clustering them in common latent space.

A method to identify opinion features from online reviews by exploiting differences in opinion feature statistics across two corpora, one domain-specific corpus and the other domain-independent corpus was proposed by Hai *et al*.,[12]. A candidate opinion features list was extracted from a domain review corpus by defining a syntactic dependency rules set. For every extracted candidate feature, it'sIntrinsic-Domain Relevance (IDR) and Extrinsic-Domain Relevance (EDR) scores are found out on domain-dependent and domain-independent corpora, respectively. Experiments on 2 real-world review domains show

that the new IEDR approach outperforms many methods in identifying opinion features.

Sentiment classification is an important OM sub-problem, which classifies net news comments as positive or negative to help government identify netizens' viewpoints. Most sentiment classification research uses only use single classifier like kNN, Naive Bayes or SVM. Two multiple classifiers integration algorithms like Bagging and Boosting were used by Fan *et al.*,[13], for sentiment classification. Various feature selection methods are investigated. Experiments show that AdaBoost approach, a Boosting method, usually achieved better performance than Bagging and statistic based single classifier and feature selection was better than POS-based method for sentiment classification of Chinese net news comments.

## METHODOLOGY

The obtained blog data from various medical information sites is highly noisy. The first step was to remove unwanted data by using stop words and stemming. The Term Document Frequency – Inverse Document Frequency is computed to obtain the weight vector of the terms used in all the document. To reduce dimensionality, PCA and Kernel PCA are used, and the best features selected using proposed Fish Swarm Optimization algorithm.

**Term Document Frequency (TDF)**

The TDF can be computed as a set of documents "x"and a set of terms "a" so that each document is modelled as a vector *v* in "a"with the dimensional space $_{R^a}$. The term frequency can be denoted as $freq(x,a)$ which expresses the number of occurrence of the term "a" in document "x". The term-frequency matrix $TF(x,a)$ measures the term association "a" to a given document "x". $TF(x,a)$ is assigned as zero when the document contains no term in it and $TF(x,a)=1$ when term "a" occurs in the document "x" or while using a relative term frequency in which the term frequency is against the total occurrences of all document terms.Frequency is normalized as[14]:

$$TF(x,a)=\begin{cases}0 & freq(x,a)=0\\ 1+\log\left(1+\log\left(freq(x,a)\right)\right) & otherwise\end{cases} \quad ...(1)$$

**Term Frequency - Inverse Document Frequency (TF-IDF)**

IDF is a measure of word's importance and appears in many heuristic measures in information retrieval. Till now IDF has been a popular and heuristic measure of a word's importance and defined as the logarithm of the ratio of documents number,which is having a given word. So it automatically ensures that the rare words have high IDF and common words like "the" have only low IDF, which measures a word's ability to discriminate between documents. Text Classification is used to assign a text document to a pre-defined class set automatically, through the machine learning. Classification is on the basis of significant words/key-features of a text document. When the classes are pre-defined,then it becomes a supervised machine learning task.

$IDF(a)$

IDF is used to represent the scaling factor. When a term "*a*" occurs frequently in many documents, then its importance is then scaled down because of its lowered discriminative power. The

can be defined as follows:

$$IDF(a)=\log\frac{1+|x|}{x_a} \quad ...(2)$$

$x_a$ is the set of documents,which consists of term "*a*".

TDF-IDF uses text categorisation metric having two scores such as term frequency and

**Table 1.** Classification Accuracy

| Classification Accuracy % | Naïve Bayes | CART | K NN |
|---|---|---|---|
| Without Feature Selection | 0.7422 | 0.7733 | 0.7378 |
| Feature transformation using PCA | 0.7556 | 0.7844 | 0.7467 |
| Feature transformation using KPCA | 0.7627 | 0.7911 | 0.7578 |
| Feature transformation using PCA and selection using FSO | 0.7756 | 0.8022 | 0.7911 |
| Feature transformation using KPCA and selection using FSO | 0.7844 | 0.8244 | 0.8178 |

inverse document frequency. Term frequency is used to count the times a term occurs in a document while inverse document-frequency is attained by dividing total documents by documents where a specific word appears repeatedly. The multiplication of values results in a high score for frequently occurring words in limited documents. A low score is intended for terms appearing frequently in all documents.

**Principal Component Analysis (PCA)**

Component analysis aims to locate "right" features from data without supervision[15]. PCA lowers feature dimensional space and is a technique used to reduce multivariate datasets[16] dimensionality. PCA analyzes data table of observations by several dependent variables, which are inter-correlated. Its aim is extracting important information from data table and expressing such information as new orthogonal variables set called principal components. PCA represents observations and variables similarity pattern by displaying them as map points[17].

PCA lowers original feature space's dimensionality linear subspace where new features have largest variance. This is *dimensionality reduction*, as vector $\bar{x}$ with original data andN-dimensional is lowered to a compressed vector that is M-dimensional, where M<N.

A vector is coded into a vector with reduced dimension. Vector is stored, transmitted or processed resulting in vector, which can be decoded back to a vector. The last vector is due to approximation which is reached by storing, transmitting/processing vector[18].

Diagram's encoder must perform a linear operation, using a matrix:

$$\bar{c} = \overline{\overline{Q}}\,\bar{x} \quad \text{...(3)}$$

Decoder is also a linear operation, written as vector elements of sum multiplied by matrix columns

$$\overline{\overline{Q}}: \quad \tilde{\bar{x}} = \bar{c}^T \overline{\overline{Q}}^T \rightarrow \tilde{\bar{x}} = \sum_{i=1}^{M} c_i \bar{q}_i \quad \text{....(4)}$$

**Kernel PCA**

The PCA's nonlinear version is constructed by using kernel function known as Kernel Principal Component Analysis (KPCA).

KPCA aims to map data points from input space to high-dimensional (inûnite-dimensional) feature space *f*

$$\Phi = \mathbb{R}^n \rightarrow \mathcal{F} \quad \text{...(5)}$$

and perform PCA in F. The space F and also mapping Ö could be complicated [19]. But using the so-called kernel trick, it avoids using Ö: PCA in F is formulated to ensure that only F's inner product is needed which is a nonlinear kernel function

$$\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \quad \text{...(6)}$$

**Table 2.** Precision and Recall obtained

| | Naïve Bayes | | CART | | K NN | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| **Without Feature Selection** | | | | | | |
| Positive opinion | 0.73 | 0.7689 | 0.7566 | 0.76 | 0.7489 | 0.7422 |
| Negative opinion | 0.7559 | 0.7156 | 0.7589 | 0.7556 | 0.7445 | 0.7511 |
| **Feature transformation using PCA** | | | | | | |
| Positive opinion | 0.7421 | 0.7289 | 0.7662 | 0.7867 | 0.7533 | 0.76 |
| Negative opinion | 0.7336 | 0.7467 | 0.7808 | 0.76 | 0.7578 | 0.7511 |
| **Feature transformation using KPCA** | | | | | | |
| Positive opinion | 0.7583 | 0.8089 | 0.7787 | 0.8444 | 0.7712 | 0.8089 |
| Negative opinion | 0.7952 | 0.7422 | 0.8301 | 0.76 | 0.7991 | 0.76 |
| **Feature transformation using PCA and selection using FSO** | | | | | | |
| Positive opinion | 0.7735 | 0.8044 | 0.7992 | 0.8489 | 0.7824 | 0.8248 |
| Negative opinion | 0.7963 | 0.7644 | 0.8389 | 0.7867 | 0.8175 | 0.7544 |
| **Feature transformation using KPCA and selection using FSO** | | | | | | |
| Positive opinion | 0.7764 | 0.8178 | 0.8241 | 0.8247 | 0.8017 | 0.8622 |
| Negative opinion | 0.8075 | 0.7644 | 0.7931 | 9.8362 | 0.851 | 0.7867 |

$$(x, y) \rightarrow k(x, y), \quad \text{...(7)}$$

This calculates every vector pair's real number from input space

**Fish Swarm Optimization**

FSO's idea is simulating fish swarm foraging. Fish swarm finds nutrient-rich areas in oceans and lakes quickly. When fish swarms behavioral characteristics are observed, it features clustering, following and foraging. Following are typical fish swarm behaviour:

**Foraging behavior**

Fish swim freely and randomly. When food is located, they quickly swim to the spot where food is increased[20].In general, the fish stroll at random. When the fish is discovered in a water area with more food, then they will go quickly towards that area. This behaviour can be expressed as follows:

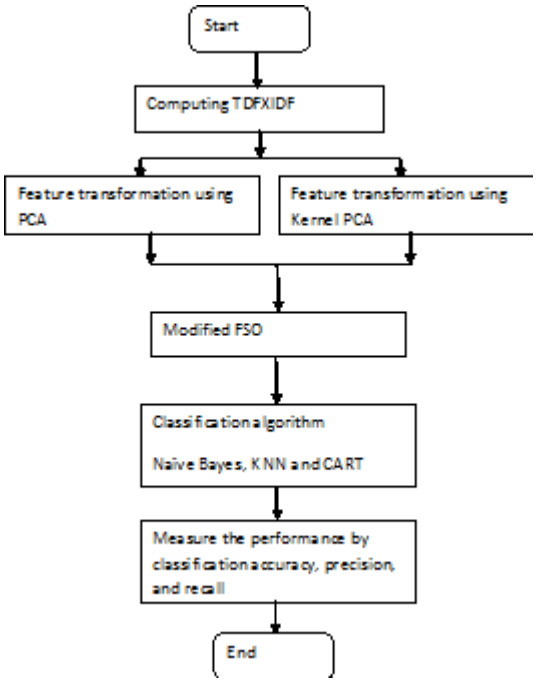


**Fig. 1.** Flowchart for Proposed Methodology

$$X_j = X_i + visual.rand() \quad \text{...(8)}$$

$$X_i = X_i + step.(X_i - X_j)/ \| X_i - X_j \| \text{ if } Y_j > Y_i \quad \text{...(9)}$$

$$X_i^{(t+1)} = X_i^{(t)} + step.rand() \text{ else} \quad \text{...(10)}$$

Where is the present position of the fish,is the position of the fish, which has discovered the food[21].

**Clustering behavior**

When a fish swarm moves, they cluster as a group to survive and to avoid harm. Fish follow three rules in clustering:

- Separation rule. They avoid overcrowding and keep distance from each other.
- Alignment rules. They move in the same direction with partners as much as possible.
- Cohesion rule. They move to the center of partners.

**Following behavior**

When one/several fish find food, other fish nearby swim in that direction.

FSO is applicable to a range of practical project optimizations. Compared to traditional optimization algorithms and evolutionary computations, FSO characteristic is summarized as follows:

(i) FSO has faster convergence rate and can solve problems with real-time control requirement.

(ii) For some optimization problems without needing precise solutions, FSO provides a feasible solution.

The steps for fish swarm optimization are as follows:

1. Generating the initialswarm of feature subset.
2. The position of individual is calculated.
3. Every individual chooses one of the behavior operations to perform as iteration in population. The behaviour operations are such as Foraging, Clustering and Following are performed.
4. The step 3 gets repeated for all available individuals.
5. With the iteration of population, the best individual is decoded among calculated individual[22].


**Fig. 2.** Process of PCA

**Classifiers used**

The proposed feature techniques were evaluated with a probability based classifier, distance based classifier and a tree based classifier.

**Naive Bayes classifier**

Naive Bayes classifier is based on strong (Naïve) Bayes theorem with strong independence assumption and suits cases with high input dimensions. The Naive assumes that features are conditionality independent, for example, in a document word occurrences (features) do not depend on each other. Naïve Bayesian is a statistical classifier performing probabilistic prediction, i.e. predicting class membership probabilities.

Following are Naïve Bayes algorithms:

$$P(C_i|V) = \frac{P(V|C_i)P(C_i)}{P(V)} \quad ...(11)$$

Where is a document represented in n-dimensional attribute vector and $C_1,.....,C_m$ represents m class. But it is computationally expensive to compute $P(V|C_i)$. To reduce computation, naïve conditional independence assumption of class is made. Thus,

$V = (v_1,.....,v_n)$

$$P(V|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad ...(12)$$

**K-Nearest Neighbour Classification**

KNN is an important non-parameter, supervised learning algorithm. Classification rules are generated by training samples without any added data. KNN classification algorithm predicts test sample's category according to K training samples, which are nearest neighbors to test sample and judge it to that category, which has largest category probability.

Adding weights refines class labels. Tuning *k* gets higher accuracy. k Nearest Neighbour method canbe easily understood and implemented[23].

$$p(x) \cong \frac{k}{NV} \quad ...(13)$$

Similarly, probability density function $p(\mathbf{x}|\mathbf{Hi})$ of observation **x** conditioned to hypothesis is approximated. Ifit is assumedis number of patterns associated to hypothesis

$$H_i, i = 1...C, \text{ so that } N1 + ... + NC = N. \quad ...(14)$$

**Classification and Regression Trees (CART)**

Classification and Regression Trees (CART) handle numerical/categorical variables. CART's advantages are its robustness to outliers. Usually splitting an algorithm isolates individual node/nodes outliers. A CART's practical property is it its classiûcation/regression trees structure being invariant regarding independent variables monotone transformations. A variable is replaced by its logarithm or square root value without its tree structure changing[24].

$$i(t) - p_L i(t_L) - p_R i(t_R). \quad ...(15)$$

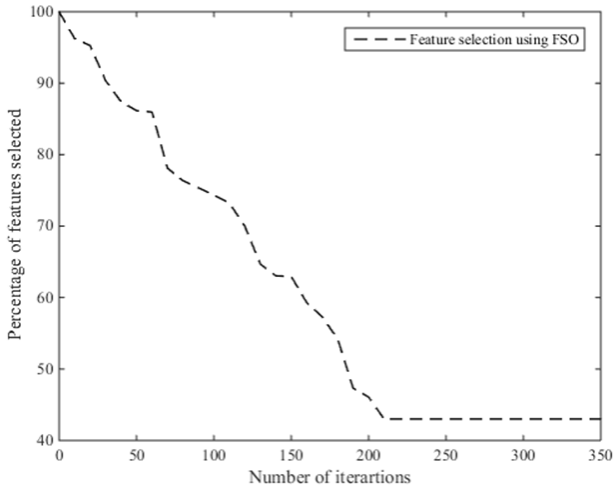


**Fig. 3.** The convergence characteristics of the selected features

## RESULTS AND DISCUSSION

450 opinions for insomnia and common cold were collected with opinions on various treatment techniques. Of the 450 opinions, 368 opinions were positive opinion with a treatment showing successful recovery and 82 opinions being negative with poor recovery.Experiments were conducted under the scenarios listed.

i. Feature transformation using PCA(PCA)
ii. Feature transformation using Kernel PCA(KPCA) and
iii. Feature transformation using Kernel PCA with feature selection using FSO (PCA-FSO)

Table 1 shows the classification accuracy various scenarios.

From Table 1 it is observed that the CART achieves better classification accuracy at 82.44 % compared to the other classifier used. Table 2: Precision and Recall with various techniques

For Naïve Bayes KPCA FSO achieves better precision than other techniques by an average of 1.08% whereas in CART the improvement is on an average of 2.73% for the proposed feature selection. Similarly, KNN showed improvements of 3.42. Similarly,improvements in the recall are in the tune of 1.15%, 2.77%, and 3.38 % respectively.

The convergence characteristics of the feature selected are shown in figure 3. It is observed that FSO converges within 210 iterations and reduces the number of features by almost 57 %. This improves the computational time, which is crucial for large datasets.

## CONCLUSION

This workinvestigated feature extraction methods/classification technique efficacy for Opinion Mining of medical blogs. The proposed framework involved obtaining opinions from various medical information blogs, removing the noise using stop words and stemming. Feature vectors were obtained using TF-IDF. The obtained feature vectors were transformed using PCA and Kernel PCA. Important features were selected using the proposed FSO.CART's classification accuracy was 82.44 % and performed better than Naïve Bayes and KNN.

## REFERENCES


1. Hajmohammadi, M. S., Ibrahim, R., & Othman, Z. A. Opinion mining and sentiment analysis: A survey. *International Journal of Computers & Technology,* 2012; **2**(3): 171-178.
2. Liu, B. Sentiment analysis and subjectivity. Handbook of natural language processing, 2010; **2**: 568.
3. Ding, X., Liu, B., & Yu, P. S.A holistic lexicon-based approach to opinion mining. In Proceedings of the international conference on Web search and web data mining, 2008(pp. 231-240
4. Buche, A., Chandak, D., & Zadgaonkar, A. Opinion Mining and Analysis: A survey, 2013; arXiv preprint arXiv:1307.3336.
5. K.Thangavel "Fuzzy - Rough Feature Selection with Ð -Membership Function for Mammogram Classification", *IJCSI International Journal of Computer Science Issues*, 2012; **9**(4): No 3.
6. Roselin, R., Thangavel, K., & Velayutham, C. Fuzzy-Rough Feature Selection for Mammogram Classification. *Journal of Electronic Science and Technology (JEST)(International),* 2011; **9**(2): 124-132.
7. Hossain, M. R., Oo, A. M. T., & Ali, A. B. M. The Effectiveness of Feature Selection Method in Solar Power Prediction. *Journal of Renewable Energy,* 2013.
8. Ohana, B., & Tierney, B. (2009, October). Sentiment classification of reviews using SentiWordNet. In 9th. IT & T Conference (p. 13).
9. Srivastava, R., Bhatia, M. P. S., Srivastava, H. K., & Sahu, C. P. (2010, September). Exploiting grammatical dependencies for fine-grained opinion mining. In Computer and Communication Technology (ICCCT), 2010 International Conference on (pp. 768-775). IEEE.
10. Liu, L., Lv, Z., & Wang, H. (2012, October). Opinion mining based on feature-level. In Image and Signal Processing (CISP), 2012 5th International Congress on (pp. 1596-1600). IEEE.
11. Pan, S. J., Ni, X., Sun, J. T., Yang, Q., & Chen, Z. (2010, April). Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the 19th international conference on World wide web (pp. 751-760). ACM.
12. Hai, Z., Chang, K., Kim, J., & Yang, C. (2013). Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance.

13. Fan, W., Sun, S., & Song, G. (2011, April). Sentiment Classification for Chinese Netnews Comments Based on Multiple Classifiers Integration. InComputational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on (pp. 829-834). IEEE.
14. Liu, K. L., Meng, W., Qiu, J., Yu, C., Raghavan, V., Wu, Z., ... & Zhao, H. (2007, June). AllInOneNews: development and evaluation of a large-scale news metasearch engine. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data (pp. 1017-1028). ACM.
15. Timotius, I. K., Setyawan, I., & Febrianto, A. A. Face recognition between two Person using kernel principal component analysis and support vector machines. *International Journal on Electrical Engineering and Informatics*, 2010; **2**(1): 55-63.
16. Langley, P., Bowers, E. J., & Murray, A. Principal component analysis as a tool for analysing beat-to-beat changes in electrocardiogram features: application to electrocardiogram derived respiration. *IEEE Trans. Bio. Eng*, 2010; **57**(4): 821-829.
17. Abdi, H., & Williams, L. J. Principal component analysis. Wiley Interdisciplinary Reviews: *Computational Statistics*, 2010; **2**(4): 433-459.
18. Jolliffe, I. (2005). Principal component analysis. John Wiley & Sons, Ltd.
19. Honkela, A., Harmeling, S., Lundqvist, L., & Valpola, H. (2004). Using kernel PCA for initialisation of variational Bayesian nonlinear blind source separation method. In Independent Component Analysis and Blind Signal Separation (pp. 790-797). Springer Berlin Heidelberg.
20. Krishnamoorthy, S., Saple, A. K., & Achutharao, P. H. (2008, January). An integrated query optimization system for data grids. In Proceedings of the 1st Bangalore Annual Compute Conference (p. 17). ACM.
21. Kumar and Sahay "nonlinear system identification using a novel immune artificial fish swarm algorithm", 2012.
22. Al-Ani "Feature Subset Selection Using Ant Colony Optimization".
23. Kulkarni S, Lugosi G, Venkatesh S. (1998) "Learning Pattern Classification – A Survey", IEEE Transactions on Information Theory, Vol. 44, No. 6, October 1998.
24. Timofeev, R. (2004). Classification and regression trees (cart) theory and applications. Humboldt University, Berlin.