Opinion Mining using Hybrid Methods

K.Umamaheswari, Ph.D
Professor
Department of Information Technology
PSG College of Technology
Coimbatore-4

S.P.Rajamohana
Assistant Professor
Department of Information Technology
PSG College of Technology
Coimbatore-4

G.Aishwaryalakshmi
PG Scholar
Department of Information Technology
PSG College of Technology
Coimbatore-4

ABSTRACT
Opinion mining is opinion of the public that is given by each user about a particular product. People post many comments and messages about a movie posted in these social network. The comments of each user will be taken as opinions for each movie posted in these web forums.In this paper the rating of movie in twitter is taken to review a movie by using opinion mining This paper proposed a hybrid methodsusing SVM and PSO to classify the user opinions as positive, negative for the movie review dataset which could be used for better decisions.

Keywords
Opinion Mining, Feature Extraction, PSO, SVM,

1. INTRODUCTION
Opinion mining allows information on what people are saying about products and take action immediately in relation to negative feedback in their business. Social networks are very popularbecause people use internet frequently now a days. Generally people will share their opinion, attitudes areconnected through social medias like Facebook, Twitter and linkedin. people create messages called tweets in twitter microblogs. More than 340 + million peoplerate different movies posted in twitter blog on each day.

Opinion mining research were started with the identification of opinion related with words, such as good, better, worst and bad (i.e., bothpositive and negative comments). It also involves the identification of objectivity in statements. Opinion mining is very useful for clients ,organizations as well to evaluate the opinions and towards their corporation and its product, the reviews from organization can given about product straight from people in social networking such as Tweets, comments and short messages. Many opinion mining methods were released to determine feedback given by the people to be positive or negative[5].

These methods released based on real time scenarios like movie reviews, which lead to a commercial success of this research.The top product based companies arebased on this opinion mining methods.The recent trend is Opinion mining in text mining [TM] with respect to the objectivity analysis [3].

SupportVectorMachine (SVM) is a novel machine learning technique depending on the statistical learning concept, it resolves the over-fitting, local optimal solution and has outstanding generalization capability in the scenario of minor sample. On the other hand SVM is impacted due to the problems of selecting suitable parameters. Particle Swarm Optimization (PSO) is an optimization technique is simple to apply and there are few parameters to modify [3]. The proposed work is to classify sentiment of movie reviews from twitter data by hybrid methods of PSO and SVM.

2. LITERATURE SURVEY
2.1 Existing System
SMO SVM: Social Media Optimization(SMO)SVMis used to solve quadratic programming (QP) problem that arises during the training of support vector machines. SMO is broadly used for training SVM and is implemented by LIBSVM tool.

It is an iterative algorithm to solve an optimization problem described above. SMO breaks the problem into subproblems, then it can be solved using analytical method. Because of the linear equality constraint involving the Lagrange multipliers α_i, the smallest possible problem involves two such multipliers.

Chi-Squared Test:A chi-squared testis a statisticalhypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. A chi-squared test is a test in which this is *asymptotically* true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. The chi-square (I) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.The difference between the expected and observed is analysed to find whether it is completed based on sampling variation or its real difference.Pearson's chi-squared test and Yates's correction for continuity introduce some error. This reduces the chi-squared value obtained and thus increases its p-value.

Naive Bayes Classifier: Navie Bayes-classifier is a probabilistic classifier based on conditional probability. The probability for each data to be classified into positive or negative is calculated based on the Bayes theorem. Bayesian techniques use mathematical formulae in order to analyze the content of the message. In its common form Bayes theorem is depicted in the equation 2.1

$$P\left(\frac{A}{B}\right)=\frac{P\left(\frac{A}{B}\right).P(A)}{P(B)} \dots\dots\dots\dots\dots\dots\dots\dots\dots (2.1)$$

Bayesian classifier works on the events which are dependent and the probability of an event occurring in the future that can be detected from the previous occurring of the same event[11]. Naïve bayes classifier technique has become a very popular method filtering techniques.

K-Nearest Neighbors Classifier: KNN techniques is mapped to its feature and similarity and K-nearesttraining documents

are measured. For each class of documents Rough Sets Classifier Method: Rough set has a great ability to compute the reductions of information systems.

Artificial Neural Network: These networks are computational models inspired bybiological neural networksisdepend on a large number of inputs and are generally unknown used to estimate functions ANN are generally presented as interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition.

Limitations

- Time complexity of atleast O($n^2 \log n$) is required, Sensitivity to noise and outliers.
- Breaking large clusters.
- No objective function is directly minimized.

These limitations are addressed in the proposed method.

3. FRAME WORK MODEL

The opinion mining systems performs data collection,preprocessing,featureselection,machine learning and classification.The frame work of the system is depicted in the Fig 3.1

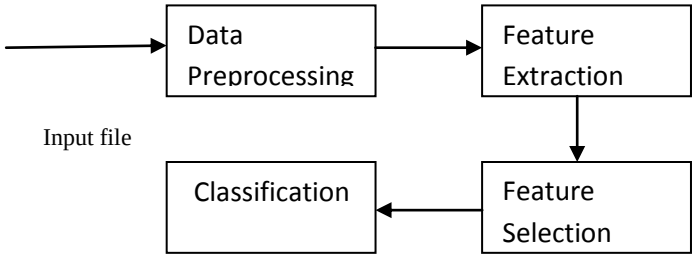

Fig 3.1 Framework Model

4. PROPOSED SYSTEM

The proposed Particle Swarm Optimization method is described.

4.1 Data Preprocessing

The movie dataset consisting of 3000 user data with their user id,rating and which genre they belong toThen we have processed data using filter data,data cleansing and export to csv file and elimination of redundant data. The result of 200 data with the user id,genre and their rating are obtained.

Filtering The Data:The data is combined together into movie id,movie name, genre and rating using clustering.

Data Cleansing **:**The redundant names of movie are eliminated**.**For eg if movie id="7264,12342" and if the rating made by two or more users are same for the particular movie then only one rating will be displayed. Hence it avoids redundant ratings.

Imported To CSV File **:**The preproccessed data to be imported into a CSV file after all the repeated data has been removed.

4.2 Feature Extraction

The features are extracted using the proposed method. The algorithm of PSO is stated below. Each particle will modify its current position and velocity according to the distance between its current position and pbest, and the distance between its current position and gbest.

```
For each particle
{
 Initialize particle with feasible random no
}
        Do
        For each particle
                {
                 Calculate the fitness value
                fitness value is better than the best fitness
                value  then Set current value as the new p
                best
                 }
        particle should be chosen with the best fitness value
        of all the particles as the gbest
          For each particle
        {
                particle velocity  is calculated based on
                velocity update equation
                Update particle position based on position
                update equation
        }
While minimum error criteria is not attained
```

4.3Machine Learning and Classification

The classification problem could be restricted to two-class problem without loss of generality. The goal is to separate the two classes by a function which is induced from available examples. The ultimategoal is to produce a classifier that will work well on unseen examples, i.e. it generalizes well. Generallyby using many linear classifiers the data can be separated, but the only one problem that maximizes the margin is machine learning problem. Linear classifier is termed the optimal separating hyper plane. SVMis the most accurate method of classification and it has the highest precision in delivering the results.

5. PERFORMANCE MEASURE

Performance analysis are based on the evaluation measures such as precision, accuracy and recall.

Evaluation Measure:

Precision*:* precision (p) is the number of relevant documents identified in movie, it is evaluated usingEq (5.1)

$$precision = \frac{|\{relevantdoc\} \cap \{retriveddoc\}|}{|\{retriveddoc\}|} \dots (5.1)$$

Recall:Recall (R) is the percentage of all data that are correctly classified as positive and negative using the Eq(5.2).

$$recall = \frac{|\{relevant\ doc\} \cap \{retrived\ doc\}|}{|\{relavant\ doc\}|} \dots\dots\dots - (5.2)$$

Accuracy:Accuracy (A) is the percentage of all the datathat are correctly categorized. It is calculated using the Eq(5.3).

$$F_1 = 2.\frac{precision.recall}{precision + recall} \ldots\ldots\ldots\ldots (5.3)$$

The performance measure is compared with SVM and hybrid method

	SVM+PSO	SVM
Precision	78.4211	71.4286
Recall	76.2696	80.1727
Accuracy	81.6087	67.6734

6. RESULT ANALYSIS

The comparison result of SVM and SVM-PSO (after datacleansing).The phases of the proposed system are tabulated in Tab 6.2.

6.1 Data Preprocessing

The preprocessed data is tabulated in Tab6.1.

Tab 6.1 Results of Data Preprocessing (sample features)

User id	Movie name	comedy	short
7264	the rink (1916)	1	1
12349	the kid (1921)	1	0
14872	entracte (1924)	0	1
19729	the broadway melody	0	0

6.2 Result of SVM and PSO

The result of SVM and PSO ispresented in Tab6.2

Tab 6.2 Result of SVM and PSO

Col 1	Col 2	Col 3
0.1874	0.7168	0.0762
0.8620	0.1548	0.8540
0.7363	0.0496	0.2921
0.0926	0.5247	0.7489
0.7948	0.5787	0.9955

Figure 6.1 describes that the comparison result of SVM and SVM-PSO with data cleansing. Based on Figure 6.1, SVM-PSO gives better solution for accuracy and precision. On the other hand, SVM provides better value for recall than SVM-PSO.The best accuracy level that gives in this study is 81.60% thatbeen achieved by SVM-PSO after data cleansing.

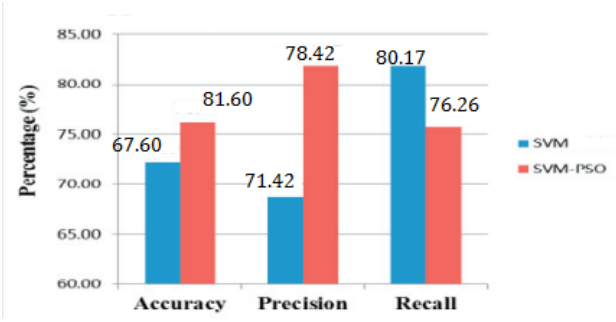

Fig 6.1 Comparison Result of SVM and SVM-PSO

7. CONCLUSION

The proposed system provides an efficient method to get a better decision about the movies for the customer. This paper uses the hybrid method and has improved accuracy compared to individual SVM after the hybridization of SVM-PSO.The accuracy level of SVM-PSO still can be improved using enhancements of SVM that might be using SVM with other optimization method. The work done in this research is only related to classification opinions into two classes, positive and negative class. The future work, a multiclass of sentiment classification such as positive, negative and neutral can be considered.

8. REFERENCES

[1] X. Yu, Y. Liu,X.Huang and A.An, "Mining online Reviews for Predicting Sales Performance", IEEE Transaction on Knowledge and Data Engineering, vol. 24, No. 4, pp. 720-734, Apr 2012.

[2] A.Pak and P.paroubek,"Twitter as a Corpus for Sentiment Analysis and Opinion Mining",Proceedings of the seventh international conference on language resources and evaluation (LREC' 10),pp.1320-1326,nov 2010.

[3] Moontae lee, Patrick Grafe, "Multiclass Sentimental Analysis with Restaurant Reviews",Department of Computer Science, Stanford University,June 2010.

[4] B. Liu, Web Data Mining, Second Edi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[5] S.F. Pratama, A.K. Muda, Y.H.Choo and N.A. Muda, "PSO and Computational Inexpensive Sequential Forward Floating Selection in Acquiring Significant Features for Handwritten Authorship," in 2011 11th International Conference on Hybrid Intelligent Systems (HIS), 2011, pp. 358-363.

[6] S. L. Salzberg. "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach",Data Mining and Knowledge Discovery, vol. 1, pp. 317 -328, 1997..

[7] Ethem Alpaydin. 2004. Introduction to Machine Learning(Adaptive Computation andMachine Learning).The MIT Press. Hayter Anthony J. 2007. Probability and Statistics for Engineersand Scientists. Duxbury, Belmont, CA, USA. Kushal Dave, Steve Lawrence, and David M. Pennock. 2003.

[8] Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.,2009

[9] J. Bollen, H. Mao, and X. Zeng ," twitter mood predicts stock market ",Journal of Computational Science, vol. 2, no. 1, pp. 1-8, Mar. 2011.

[10.] x.yu y.liu ,x.haung and A,AN ," mining online reviews for predicting sales performance", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 4, pp. 720-734, Apr. 2012.

[11] Umamaheswari K., Sumathi S., Aparna V. and Arthi A., 'Text Classification using Enhanced Naïve Bayes with Genetic algorithm', International Journal of Computer Applications in Engineering, Technology and Sciences, Gujarat, Vol. 1, No. 2, pp. 263-270, 2009.