DATA CLASSIFICATION USING MODIFIED VERSION OF SUPPORT VECTOR MACHINE

Krina Vasa[1], Arindam Chudhuri[2], Sanjay Bhanderi[3]

[1] *PG Student, Computer Engineering, MEFGI, Gujarat, India*
[2] *Associate Professor, Computer Engineering, MEFGI, Gujarat, India*
[3] *Assistance Professor, Computer Engineering, MEFGI, Gujarat, India*

ABSTRACT

Text classification is a process of categorized data according to their class. With the instant growth of information, text classification has become the vital techniques for handling and organizing text data. In general, text classification plays an important role in information extraction and summarization, text retrieval, and question-answering such as medical diagnosis, news group filtering, spam filtering, and sentiment analysis. The process of classification consists four stages: text preprocessing, feature extraction, training classifier and training model. In this first stage the dataset is divided into training data and testing data. After that data is preprocessed and features are extracted from that data then after classification model is constructed. The data is classified using machine learning techniques and statistical techniques such as k-nearest neighbors, support vector machine, naive Bayesian method. The classification task also exemplifies the hybrid approach of text classification techniques. In this research we provide a modified version of support vector machine with better membership function for text classification of text data. In that the noisy and inherent data is handled by fuzzy support vector machine and its fuzzy membership function which is used hyperbolic tangent kernel.

Keyword : *Data Classification, Support Vector Machine, Fuzzy Support Vector Machine.*

1. INTRODUCTION

Text classification is the performance of separating a set of input documents into two or more classes where each document can be said to belong to one or multiple classes. In classification systems, groups of words or terms are collected together and organized. Each of these terms will be associated with a particular concept. Systems of classification have classically been hierarchical, that means more detail is gained the extra down the hierarchy one proceeds, while concepts are linked and planned around mutual characteristics. The documents to be classified may be texts, images, music, etc. The problem of classification has been broadly studied in the data mining, machine earning, information retrieval and database with applications such as email routing, sentiment analysis, spam filtering, target marketing, and language identification.

A Support Vector Machine (SVM) is a classifier properly defined by a separating hyperplane. Given a set of training samples, each marked for belonging to one of two categories. An SVM model is a representation of the samples as points in space, mapped so that the samples of the separate categories are divided by a clear gap that is as wide as possible. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using the kernel trick, mapping their inputs into high-dimensional feature spaces. Two

applications where SVM exceeded other methods are the prediction of electrical charges and optical character recognition. SVM are more widely used nonparametric technique and give fair results. Exercise of SVM classification concludes hyperplane in the space as possible to maximize the distance from data points hyperplane. It is essentially the resolution of a problem of quadratic optimization. The solution of convex problems for SVM strictly is unique and global. SVM implements minimization of structural risks (SRM) principle that has high performance of generalization. As the complexity increases as a function of the number of vectors of support, SVM is built by the intermediary of trading off decreasing number of errors of training and increasing the risk of more of the adjustment of the data. However, data dependent SRM for is not rigorously SVM support the argument that the good performance of generalization of SVM is attributable to SRM [49]. Since SVM capture the geometrical characteristics of the feature space without drifting networks from the masses of data on training, it is able to extract an optimal solution with small size of the training set.

A support vector machine (SVM) learns the decision surface from two distinct classes of the input points. In many applications, each input point may not be fully assigned to one of these two classes. If we apply a fuzzy membership to each input point and reformulate the SVMs such that different input points can make different contributions to the learning of decision surface. Then it is called as fuzzy SVMs. In fuzzy SVM, membership in terms of probability is also determined for each sample to be fall in each class.

2. RELATED WORK

With the instant growth of information, text classification has become a vital technique for handling and organizing text data. Text classification plays an important part in information extraction, text retrieval, text summarization, news group filtering medical diagnosis, spam filtering, and sentiment analysis. text classification process using machine learning techniques and statistical techniques such as k-nearest neighbors, support vector machine, naive Bayesian method Decision tree, Rule-based classification, Neural network classification.

In 2014, A. Chaudhuri et al. [5] developed a novel fuzzy support vector machine (FSVM) tool or a variant of FSVM called modified fuzzy support vector machine (MFSVM). This variant is to classify the credit approval problem. In FSVM, each sample is given a fuzzy membership which denotes the attitude of corresponding point toward one class. The membership function which is a hyperbolic tangent kernel grips the impreciseness in training samples. In MFSVM, the victory of the classification lies in proper selection of the fuzzy membership function which is a function of center and radius of each class in feature space and is represented with kernel. The kernel used in MFSVM is hyperbolic tangent kernel. This kernel allows lower computational cost and higher rate of optimistic eigenvalues of kernel matrix which eases several limitations of other kernels.

In 2012, H. Tao et al. [19] projected Insurance Fraud Identification Research Based on Fuzzy Support Vector Machine with Dual Membership. For "overlap" problem in insurance fraud samples, this paper constructs the fuzzy support vector machine model with dual membership, which assigns each insurance fraud sample with dual membership by its relativity to the distance of the two types of sample mean vector. The dual membership can characterize the probability of each insurance fraud sample belonging to two categories.

In 2002, C. Lin et al. [18] described a fuzzy support vector machine. Support vector machine learns the decision surface from two distinct classes of the input points. In many applications, each input point may not be fully assigned to one of these two classes. In this paper, the author applied a fuzzy membership to each input point and reformulate the SVMs such that different input points can make different contributions to the of decision surface.

3. PROPOSED METHOD

The proposed framework has fuzzy support vector machine algorithm and its mathematical equations. Here, the membership function is a function of center and radius of each class in feature space and is represented with kernel. The figure 1 below shows a diagrammatic view of the proposed framework along with its modules and their flow of interactions. For classification purpose the dataset is divided into training and testing dataset. The training data is used for the train the proposed system. And test data is for using test the proposed system. Here fuzzy support vector algorithm is used in proposed system.

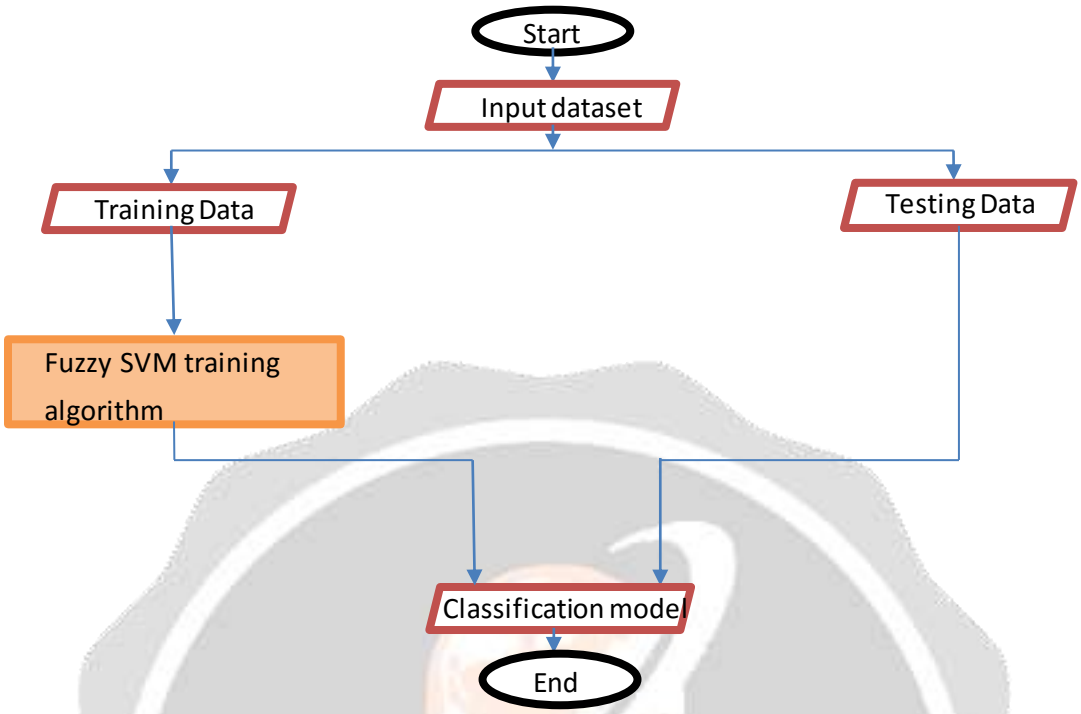

Fig 1 Proposed Framework

3.1 Mathematical Representation

In the first step the attribute Y_i spits into two classes. These equations are taken from the paper [5]. One class contains sample point $Y_i$with z_i= 1 denoted by C^+ and other class contains sample point $Y_i$with z_i= −1 denoted by C^-such that,

$$C^+ = \{Y_i | Y_i \in SP \wedge z_i = 1\}$$

$$C^- = \{Y_i | Y_i \in SP \wedge z_i = -1\}$$

C^+= positive class

C^-= negative class

SP=Sample points

Y_i= Attribute of the dataset

The hyperbolic tangent kernel is used to map sample points from input space to feature space. Kernel methods map the data into higher dimensional spaces in the hope that in this higher-dimensional space the data could become more easily separated or better structured. SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network.

The hyperbolic tangent kernel is given by,

$$K(Y_i, Y_j) = \tanh\left[\Phi(Y_i) \cdot \Phi(Y_j)\right]$$

Here $\Phi(Y_i)$ and $\Phi(Y_j)$ are kernel functions,

$$\Phi(Y_i) = \frac{\sqrt{2x_i} + \sqrt{\frac{x_i^3}{3}}}{\sqrt{2} + x_i}$$

$$\Phi(Y_j) = \frac{\sqrt{2x_j} + \sqrt{\frac{x_j^3}{3}}}{\sqrt{2} + x_j}$$

In the next step, define Φ_+ as the class of C^+ in feature space and Φ_- as the class of C^- in feature space,

$$\Phi_+ = \frac{1}{m_+} \sum_{Y_i \epsilon C^+} \phi(Y_i) f_i$$

$$\Phi_- = \frac{1}{m_-} \sum_{Y_i \epsilon C^-} \phi(Y_i) f_i$$

Φ_+ = class center of C^+
Φ_- = class center of C^-
f_i = frequency of ith sample
m_+ = number of samples of class C^+
m_- = number of samples of class C^-

Now the radius of the class C^+ and C^- are rd_+ and rd_- respectively,

$$rd_+ = \frac{1}{n} \max \parallel \Phi_+ - \phi(Y_i) \parallel$$
$$rd_- = \frac{1}{n} \max \parallel \Phi_- - \phi(Y_i) \parallel$$

Calculate the square of radius,

$$rd_+^2 = \frac{1}{n} \max \parallel \Phi(Y') - \Phi_+ \parallel^2$$
$$= \frac{1}{n} \max[\Phi^2(Y') - 2\Phi(Y') \cdot \Phi_+ + \Phi_+^2]$$
$$= \frac{1}{n} \max [\Phi^2(Y') - \frac{2}{m_+} \sum_{Y_i \epsilon c^+} \tanh[\Phi(Y_i) \cdot \Phi(Y')] + \frac{1}{m_+^2} \sum_{Y_i \epsilon C^+} \sum_{Y_j \epsilon C^+} \tanh\left[\Phi(Y_i) \cdot \Phi(Y_j)\right]]$$

$$= \frac{1}{n} \max [K(Y', Y') - \frac{2}{m_+} \sum_{Y_i \epsilon C^+} K(Y_i, Y') + \frac{1}{m_+^2} \sum_{Y_i \epsilon C^+} \sum_{Y_j \epsilon C^+} K(Y_i, Y_j)]$$

rd_+^2 = square of radius of C^+
n $= \sum_i f_i$

$$rd_-^2 = \frac{1}{n} \max [K(Y', Y') - \frac{2}{m_-} \sum_{Y_i \epsilon C^-} K(Y_i, Y') + \frac{1}{m_-^2} \sum_{Y_i \epsilon C^-} \sum_{Y_j \epsilon C^-} K(Y_i, Y_j)]$$

rd_-^2= square of radius of C^-

n $= \sum_i f_i$

Now, calculate the square of the distance between samples $Y_i \epsilon C^+$ and its class center in feature space,

$$dist_{i+}^2 = \| \Phi(Y_i) - \Phi_+ \|^2$$

$$dist_{i+}^2 = \mathrm{K}(Y_i, Y_j) - \frac{2}{m_+} \sum_{Y_j \in C^+} K(Y_i, Y_j) + \frac{1}{m_+^2} \sum_{Y_j \in C^+} \sum_{Y_k \in C^+} K(Y_j, Y_k)$$

$dist_{i+}^2$= square of distance between sample $Y_i \in C^+$ and its class center

Similarly, calculate the square of the distance between samples $Y_i \epsilon C^-$ and its class center in feature space,

$$dist_{i-}^2 = \mathrm{K}(Y_i, Y_j) - \frac{2}{m_-} \sum_{Y_j \in C^-} K(Y_i, Y_j) + \frac{1}{m_-^2} \sum_{Y_j \in C^-} \sum_{Y_k \in C^-} K(Y_j, Y_k)$$

$dist_{i-}^2$= square of distance between sample $Y_i \in C^-$ and its class center

After that, the fuzzy membership function sm_i can be defined as follows:

$$sm_i = \begin{cases} 1 - \sqrt{\frac{\|dist_{i+}^2\| - \|dist_{i+}^2\| rd_+^2 + rd_+^2}{\left(\|dist_{i+}^2\| + \|dist_{i+}^2\| rd_+^2 + rd_+^2\right) + s}} & \text{if } z_i = 1 \\ 1 - \sqrt{\frac{\|dist_{i-}^2\| - \|dist_{i-}^2\| rd_-^2 + rd_-^2}{\left(\|dist_{i-}^2\| + \|dist_{i-}^2\| rd_-^2 + rd_-^2\right) + s}} & \text{if } z_i = -1 \end{cases}$$

sm_i = Fuzzy membership function

In the proposed framework the membership function will be changed for text data because the existing membership function is not work well with noisy data and for better accuracy on different datasets the new membership functions are shown below which will implement on given dataset.

$$sm_i = \begin{cases} e^{\sqrt{\frac{\|dist_{i+}^2\| - \|dist_{i+}^2\| rd_+^2 + rd_+^2}{\left(\|dist_{i+}^2\| + \|dist_{i+}^2\| rd_+^2 + rd_+^2\right) + s}}} & \text{if } z_i = 1 \\ e^{\sqrt{\frac{\|dist_{i-}^2\| - \|dist_{i-}^2\| rd_-^2 + rd_-^2}{\left(\|dist_{i-}^2\| + \|dist_{i-}^2\| rd_-^2 + rd_-^2\right) + s}}} & \end{cases}$$

if $z_i = -1$

4. EXPERIMENTAL RESULTS

MATLAB is used to implement these equations. MATLAB stands for MATrix LABoratory and the software is built up around vectors and matrices. MATLAB had around one million users across industry and academia. This makes the software particularly useful for linear algebra but MATLAB is also a great tool for solving algebraic and differential equations and for numerical integration. Here, four parameters are used to compare with existing system and existing methods. For experiment results we used Wine dataset, Germen credit approval dataset and iris dataset.

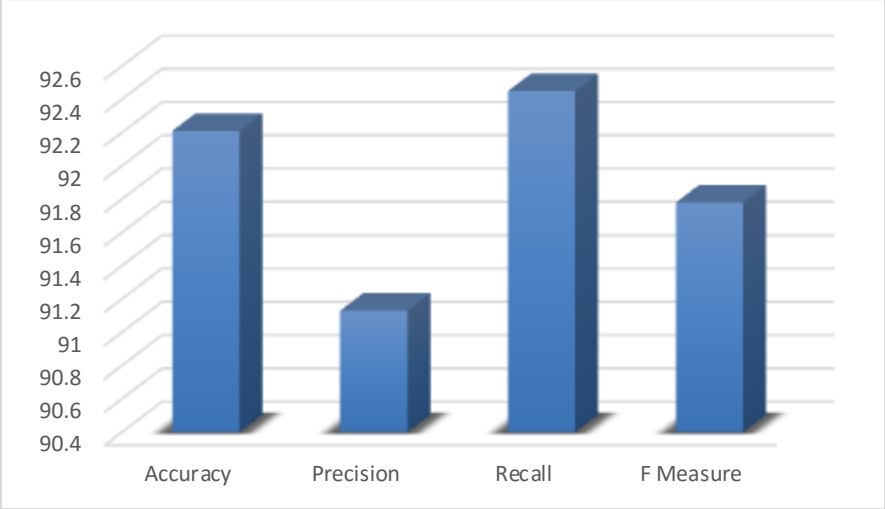

Fig 2 Performance Analysis

4.1 Comparative Analysis with Existing Solutions

The results of the proposed systemare compared with the results of existing solutions. For the same purpose the different classification methods are used on the given dataset. For comparison Support vector machine, Naive bayesian method and k-nearest neighbor classifiers are used. Following table shows comparison of proposed systemwith existing solutions on wine dataset.

	Accuracy	Precision	Recall	F Measure
Support Vector Machine	86.46	92.27	88.93	90.56
Naïve Bayes	90	90.12	89.21	89.66
k-nearest neighbor	90.11	91.11	90.76	90.93

Fuzzy Support Vector Machine	90.51	89.45	91.23	90.33
Modified Fuzzy Support Vector Machine	92.21	91.13	92.45	91.78

Table 1 Comparative analysis on Wine Dataset

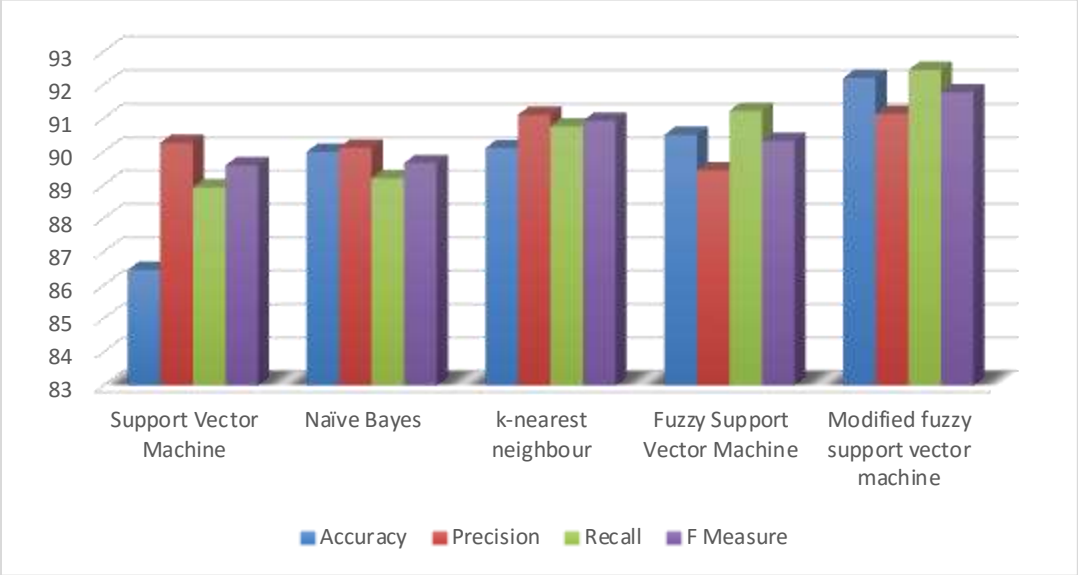

Fig 3 Comparative analysis on Wine Dataset

Following table shows comparison of proposed systemwith existing solutions on iris dataset.

	Accuracy	**Precision**	**Recall**	**F Measure**
Support Vector Machine	90.89	89.78	90.11	89.94
Naïve Bayes	82.89	87.67	77.5	82.27
k-nearest neighbor	89.67	90.33	89.5	89.91
Modified Fuzzy Support Vector Machine	91.04	88.63	90.45	89.53

Table 2 Comparative analysis on iris Dataset

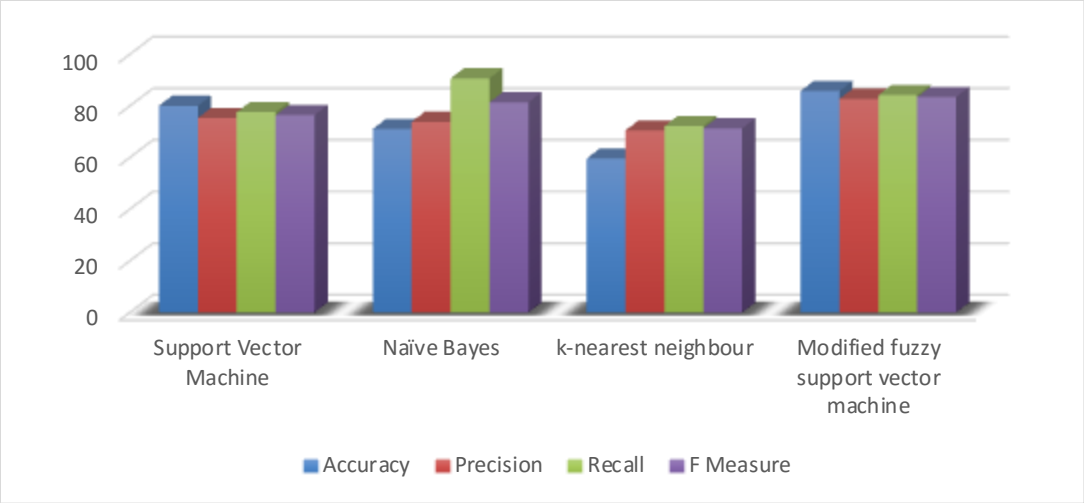

Fig 4 Comparative analysis on iris Dataset

Following table shows comparison of proposed system with existing solutions on Germen credit approval dataset.

	Accuracy	Precision	Recall	F Measure
Support Vector Machine	80.09	75.33	77.62	76.45
Naïve Bayes	71.07	73.89	90.7	81.43
k-nearest neighbor	59.67	70.55	72.25	71.38
Modified Fuzzy Support Vector Machine	85.78	82.76	84.22	83.48

Table 3 Comparative analysis on Germen credit approval Dataset

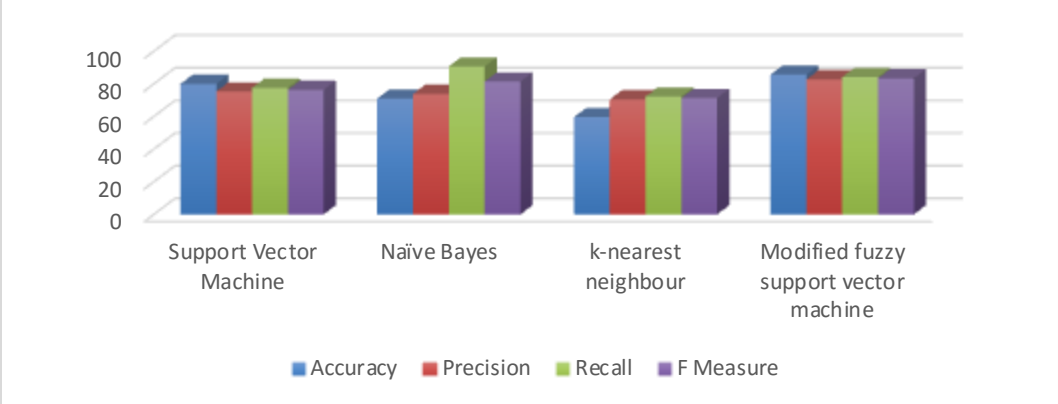

Fig 5 Comparative analysis on Germen credit approval Dataset

6. CONCLUSION

Text classification is the ultimate problem in data mining and machine learning. In today's world the demand of support vector machine is increased because of its kernel functions. But in many applications some input points are detected as outliers and may not be exactly assigned to one of the two classes. To solve these difficulty the membership function is used to give degree of probability of input point. The proposed framework will try to give better membership function and will try to implement it on various numeric dataset. The mathematical formulas for better membership function for numeric dataset were implemented. But support vector machine and fuzzy support vector machine is not work well with text or text dataset. So in future fuzzy support vector machine will implemented on text dataset classification purpose and different kernels give different results on data. So in future different kernels will also use for classification task.

7. REFERENCE

1. C. C. Aggarwal and C. Zhai, *Mining Text Data*, 2012.
2. J. Han and M. Kamber, Data Mining Concepts and Techniques, 2011.
3. M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279–289, 2015.
4. R. C. Barik and B. Naik, "A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach," Computational Intelligence in Data Mining, vol. 3, pp. 217-228, 2015.
5. R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2349–2361, 2015.
6. A. Chaudhuri, "Modified fuzzy support vector machine for credit approval classification," IOS Press and Authors, vol. 27, no. 2, pp. 189-211, 2014.
7. E. Baralis, L. Cagliero, and P. Garza, "EnBay: A novel pattern-based Bayesian classifier," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2780–2795, 2013.
8. X. Fang, "Inference-Based Naive Bayes: Turning Naïve Bayes Cost-Sensitive," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2302–2314, 2013.
9. C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," *Expert Syst. Appl.*, vol. 39, no. 15, pp. 11880–11888, 2012.
10. L. H. Lee, R. Rajkumar, and D. Isa, "Automatic folder allocation system using Bayesian- support vector machines hybrid classification approach," *Appl. Intell.*, vol. 36, no. 2, pp. 295–307, 2012.
11. M. Parchami, B. Akhtar, and M. Dezfoulian, "Persian text classification based on K- NN using wordnet," *Adv. Res. Appl. …*, vol. 7345, pp. 283–291, 2012.
12. M. Wang, H. Zhang, and R. Ding, "Research of text categorization based on SVM," *International Conference on Infomatics, Cybermetics and Computer Engineering ICCE*, Vol. 2, pp. 69-77, 2011.
13. K. Lin, and M. Chen, "On the Design and Analysis of the privacy-preserving SVM classifier," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1704-1717, 2011.
14. Z. Liu, X. Lv, K. Liu, and S. Shi, "Study on SVM compared with the other text classification methods," *2nd Int. Work. Educ. Technol. Comput. Sci. ETCS 2010*, vol. 1, pp. 219–222, 2010.
15. H. Cheng, P. Tan, and R. Jin, "Efficient algorithm for localized support vector machine," *Knowl. Data Eng. …*, vol. 22, no. 4, pp. 537–549, 2010.
16. D. Martens, B. B. Baesens, and T. Van Gestel, "Decompositional Rule Extraction from Support Vector Machines by Active Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 2, pp. 178–191, 2009.

17. W. Zhang, T. Yoshida, and X. Tang, "Text classification based on multi-word with support vector machine," *Knowledge-Based Syst.*, vol. 21, no. 8, pp. 879–886, 2008.

18. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.* vol. 2, no. 2, pp. 121–167, 1998.

19. C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *Neural Networks, IEEE Trans.*, vol. 13, no. 2, pp. 464–471, 2002.

20. H. Tao, L. Zhixin, and S. Xiaodong, "Insurance fraud identification research based on fuzzy support vector machine with dual membership," *2012 Int. Conf. Inf. Manag. Innov. Manag. Ind. Eng.*, pp. 457–460, 2012.

21. C. Moewes and R. Kruse, "On the usefulness of fuzzy SVMs and the extraction of fuzzy rules from SVMs," *Proc. 7th Conf. Eur. Soc. Fuzzy Log. Technol. LFA-2011*, vol. 17, no. July, pp. 943–948, 2011.

22. D. BOIXADER, J. JACAS, and J. RECASENS, "Upper and Lower Approximations of Fuzzy Sets," *Int. J. Gen. Syst.*, vol. 29, no. 4, pp. 555–568, 2000.