An Experiential Study of SVM and Naïve Bayes for Gender Recognization

Sneha Pradeep Vanjari
Department of Computer Engineering,
SKN-Sinhagad Institute of Technology & Science,
Lonavala,
India
Email_id: snehavanjari9900@gmail.com

Guided By
V. D. Thombre
Department Computer Engineering,
SKN-Sinhagad Institute of Technology & Science,
Lonavala, India
Email_id: thombrevd@gmail.com

Abstract: Classification and regression are the important aspects of data mining. Data mining is the systematic procedure of extracting useful data from large datasets. Naïve Bayes and SVMs are useful for classification and regression. The naïve Bayes classifier is a typical generative classifier .while The SVM classifier is a typical Discriminative classifier .The naive Bayesian (NB) classifier is one of the simple yet powerful classification methods. It is considered as one of the most effectual and significant learning algorithms for machine learning and data mining and also has been treated as a core technique in information retrieval. Support Vector Machines (SVM) are supervised learning models with associated learning algorithm that analyses data and recognize patterns, used for classification and regression analysis. Here we attempt to analyze the performance both algorithms i.e. Naïve Bayes and SVMs through the development of useful application. With the accuracy of developed application this will going to estimate the performance. This paper presents the novel idea towards the classification of the naive bayes and SVMs algorithm by analyzing the human characteristics for the sake of their gender identification.

Keywords – *Data Mining, Naive Bayesian, Support Vector Machines***.**

*****

I. Introduction:

With the rapid development of Internet applications, e-commerce and network communication, there is a geometric multiples growth for information, which has brought our lives more and more important influence. Almost all information we want can be found in the network. But what people care about most is how to dig out the most valuable information from this large quantity of information. The technology of automatic text classification is one of the basic ways to solve these problems, and it is an important research subject in information storage and retrieval. Automatic text classification has many advantages, such as needing no human intervention, saving a lot of manpower and updating quickly. Faster classification and higher precision will surely meet the practical application requirements. In the digital library, accurate and efficient text classification is the basis for good service to readers. Spam filtering, personalized news service and intelligent product recommendation which is relatively more popular at shopping sites at present are all typical applications of classification. Therefore, classification technology is more and more widely applied to real life[1],[2].

Naïve Bayes

Naive Bayes classification algorithm is one of the most effective text classification methods, and in some areas its performance can be comparable with neural networks and decision tree learning.

Advantages

1. Requires small amount of training data to estimate parameters.
2. Good results are obtained in most of case.
3. Easy to implement

Disadvantages

1. Assumption of class conditional independence leads to loss of accuracy.
2. Practically dependencies exist among variables and sometimes these dependencies cannot be modeled by naïve bayes.

Support Vector Machines (SVMs)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Advantages

1. It has a regularisation parameter, which makes the user think about avoiding over-fitting.
2. It uses the kernel trick, so we can build in expert knowledge about the problem via engineering the kernel.
3. It is an approximation to a bound on the test error rate.

II. Related Work:

For quite few year, the interesting topic for research is sentiment analysis [4]. Inorder for understanding the public's opinions and experiences, sentiment analysis technique is used for many social media applications [5]. Some of the known applications are predicting the poll ratings [6], prediction of stock market and event analyzing [3], behavioral targeting [7], etc. There are many learning methods used for analyzing the social media, and these methods are similar to the commonly used sentiment analysis methods used for reviewing products or movies. The methods can either be supervised learning methods [8] or unsupervised learning methods [6]. The social media sites generate a very large amount of data which is unlabeled, and so the unsupervised learning method becomes needy now-a-days. There is no need of training data for unsupervised learning techniques, making it efficient for using in any domain.

Lexicon-based method is one of the most characteristic ways of performing unsupervised sentiment analysis. In the existing work [4], social media is been used for prognosticating the future results. For instance, they have used one of the most favoured sites Twitter, for forecasting the box-office revenues of movie before they are released. They used a linear regression method for prediction. But this technique has limitations such as, emotion analysis not considered; implementation of NLP technique is required for improving the approach, exact entity recognition skill not used and so on. The systems that are implemented are not specific and they are used only for analyzing a single entity, thus making them insufficient for opinion analysis.

III. Implementation Details:

The fundamental focus of our framework is investigating the performance of Naïve Bayes and SVMs are used for the classification and regression. So, identification of the best algorithm is the main motivation behind this work. This identification compares the performance of these algorithms and, suggests the appropriate algorithm.

In addition to this it will be helpful in feature recognition also so that it can be used in two ways i.e. real time application and novel comparison approach via the means of real time application.

A. *System Architecture:*

System design is process of defining the architecture, modules, interface, and data for a system to gratify specified requirements.

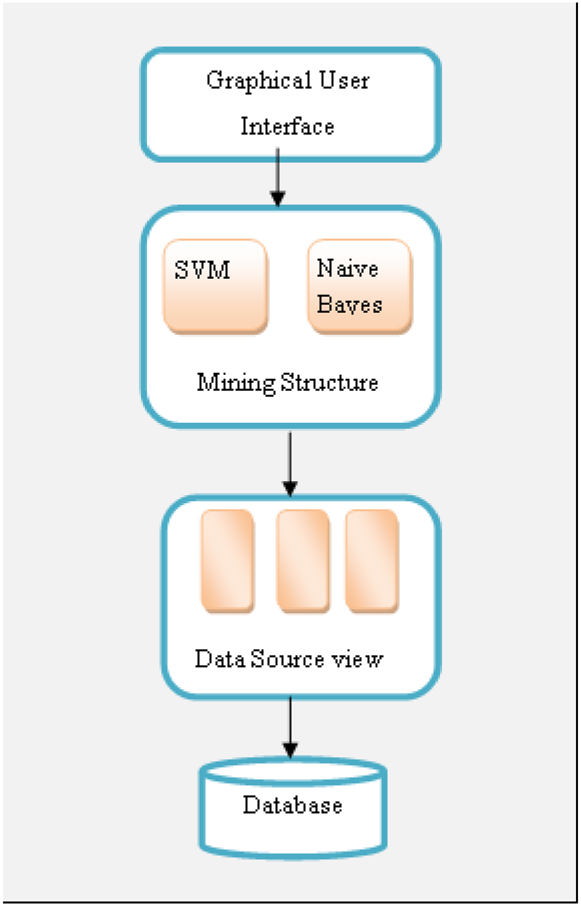

Fig.1: System Architecture

System Architecture is defined as following way:
The proposed system consist of four major components as shown in figure

1: *Graphical user interface*- Allows user to interact with the system.

2. *Mining Structure*- Uses SVM and Naive bayes processing algorithms.

3. *Data source View*- Estimates the relevant characteristics and compares it with the database.

4. *Database* - It stores the matching parameters[10].

- *Naive Bayesian Classifier*

This classifiers are probabilistic classifiers. The process based on maximum posteriori hypothesis and Bayes' theorem. They guess class membership probabilities. i.e. the probability that given test sample based on correspondence class. It assumes the class condition independence to decrease the training data requirements and cost computation[11].

- *Support Vector Machine*

It try to find out maximum margin separation hyper plane between the data of two classes to test data points by generalizing well. The basic support vector machine are discriminative classifiers and bi-nary classifiers[12].

B. Algorithm:

Input: User data (height, Weight, foot size, chest size)

Output:

The following steps explain the implementations

1. Start

2. Input given to naivesbayes algorithm I_N=*{*H_t ,W_t , F_t , C_t *}*

H_t=*Height;*

W_t=*Weight;*

F_t=*Foot size;*

C_t=*Chest size*

3. Input given to svm algorithm

I_S=*{*H_t ,W_t ,F_t , C_t *}*

H_t=*Height*

W_t=*Weight*

F_t=*Foot size*

C_t=*Chest size*

4. Take result from SVM and naives Bayes

O_n =*{*ON_m ,$ON_{f,}$ ON_o*}*

O_n = *Output of naives bayes.*

ON_m=*Output of naives bayes male.*

$ON_{f,}$=*Output of naives bayes female.*

ON_o=*Output of naives bayes other*

O_s =*{*OS_m ,$OS_{f,}$ OS_o*}*

O_s = *Output of SVM.*

OS_m = *Output of SVM male.*

$OS_{f,}$ = *Output of SVM female.*

OS_o = *Output of SVM other.*

5. Compare the result.

7. Stop

IV. Result and Experiment:

TABLE 1: Training Set

.

sex	height (feet)	weight (lbs)	foot size(inches)	chest size(inches)
Male	6	180	12	36
Male	5.92 (5'11")	190	11	37
Male	5.58 (5'7")	170	12	38
Male	5.92 (5'11")	165	10	36
Female	5	100	6	31
Female	5.5 (5'6")	150	8	33
Female	5.42 (5'5")	130	7	30
Female	5.75 (5'9")	150	9	31

The training set is used to create classifier using Gaussian distribution assumption. The examples of training set in presented in Table 1. Probability of occurrence of classes P(female) = P(male) = 0.5. This is considered frequency from the prior probability distribution in the larger population is presented in Table 3. Consider example Sex = Sample, Height = 6 Feet, Weight = 130 lbs, foot size = 8 inches. And we want to decide which posterior is greater, Male or Female

The posterior of male is gave

$$posterior(male) = \frac{p(male)p(height|male)p(weight|male)p(footsize|male)}{evidence}$$

The posterior of female is given by,

$$posterior(female) = \frac{p(female)p(height|female)p(weight|female)p(footsize|female)}{evidence}$$

TABLE 2: Prior probability distribution

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean (foot size)	Variance (foot size)	Mean (foot size)	Variance (foot size)	Mean (foot size)	Variance (foot size)
Male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01	11.25	9.1667e-01	38.25	19.1667e-01
Female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00	7.5	1.6667e+00	27.5	12.6667e+00

The evidence is calculated as,

$$evidence = p(male)p(height|male)p(weight|male)p(footsize|male) + p(female)p(height|female)p(weight|female)p(footsize|female)$$

Now we have to determine the probability distribution of Sex:

$P(male) = 0.5$

$$P(height|male) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\pi)^2}{2\sigma^2}\right) \approx 1.5789$$

where $\pi = 5.855$ and $\sigma^2 = 3.5033.10^{-2}$ are the parameters of normal distribution. It is prior defined by training set.

$$p(foot\ size|male) = 1.3112 * 10^{-3}$$

posterior numerator (male) = their product = $6.1984.\ 10^{-9}$

$$p(female) = 0.5$$

$$p(height|female) = 2.2346.\ 10^{-1}$$

$$p(weight|female) = 1.6789.\ 10^{-2}$$

$$p(foot\ size|female) = 2.8669.\ 10^{-1}$$

posterior numerator(female) = their product = $5.3778 * 10^{-4}$

Since posterior numerator is greater in the female case, we predict the sample is female.

Naive Bayes approach leads to loss of accuracy based on class conditional independence and also practical independencies are exist among variables, but it is not handled by Naive Bayes algorithm.

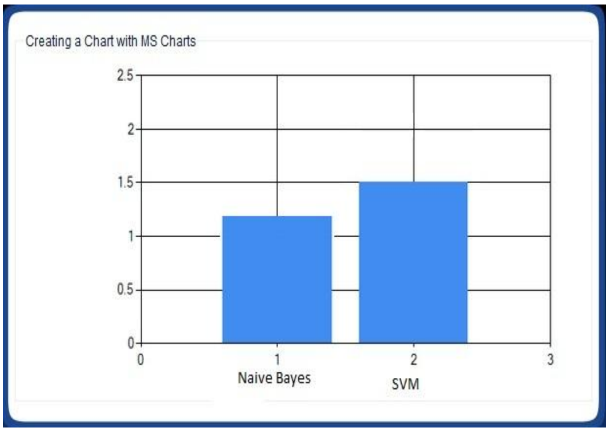

Figure 2 : Comparison between SVM and Naive Baye

Acknowledgment:

We are thankful to the teachers for their valuable guidance. We are thankful to the authorities of Savitribai Phule University of Pune and reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES:

[1] Yuguang Huang, Lei Li, “naive bayes classification algorithm based on small sample set”, Proceedings of IEEE CCIS2011

[2] K. Sankar, S. Kannan and P. Jennifer, “Prediction of Code Fault Using Naive Bayes and SVM Classifiers”, Middle-East Journal of Scientific Research 20 (1): 108-113, 2014

[3] Geor gios Paltoglou and Mike Thelwall. “Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media” ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 4, Article 66, Publication date: September 2012

[4] Xujuan Zhou,Xiaohui Tao,Jianming Yong and Zhenyu Yang. “Sentiment Analysis on Tweets for Social Events”. In Proceedings of the IEEE 17th International Conference on Computer Supported Cooperative Work in Design ,pages 557-562,2013.

[5] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. “Twitter power: Tweets as electronic word of mouth”. J. Am. Soc. Inf. Sci., 60(11):2169-2188, 2009.

[6] K. Dave, S. Lawrence, and D. Pennock. “Opinion extraction and semantic classification of product reviews”. In Proceedings of the 12th International World Wide Web Conference (WWW), pages 519-528, 2003.

[7] S. Brody and N. Elhadad. “An unsupervised aspect sentiment model for online reviews”. In HLT ’10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 804-812, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[8] B.Liu.Opinion mining. Invited contribution to Encyclopedia of Database Systems, 2008

[9] Alena Neviarouskaya, Masaki Aono, “Sentiment Word Relations with Affect, Judgment, and Appreciation”, IEEE transactions on affective computing,

[10] Roger S. Pressman, “Software Engineering: A Practitioner’s Approach”, McGRAW Hill international publication, seventh edition, 2004.

[11] R.R. Yager, “An extension of the naive bayesian classifier,” Information Sciences , vol. 176, no. 5, pp. 577–588, 2006

[12] C. Cortes and V. Vapnik, “Support-vector networks,” Machinelearning , vol. 20, no. 3, pp. 273–297, 1995