**Chapter 2**

**Review of Literature**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Sr. No.*** | ***Studies*** | ***Mining Techniques Used*** | ***Data Source*** | ***Performance***  ***(Accuracy)*** | ***Precision*** |
| 1. | *Long Sheng*  *(2011)* | *Back Propagation Neural Network* | *Movie Review* | *64%* | *60%* |
| 2. | *Rui Xia (2011)* | *Naïve Bayes, Maximum Entropy, SVM* | *Movie Review, Multi Domain Dataset* | NB- 85.8  ME- 85.4  SVM- 86.4 | - |
| 3. | *Xue Bai (2011)* | *Naïve Bayes* | *Movie*  *review* | *92%* | *-* |
| 4. | *Ziqiong (2011)* | *Naïve Bayes, SVM* | *Cantonese reviews* | *93%* | *-* |
| 5. | *Yulan He (2010)* | *Sentiment Lexicon* | *Movie Review* | *74.7%* | *-* |
| 6. | *Zhu Jian (2010)* | *Back propagation* | *Movie Review* | *86%* | *-* |
| 7. | *Melville (2009)* | *Bayesian classification* | *Blogs* | *91.2%* | *-* |

**Table 1: Literature Survey [1]**

Rui Xia (2011) made a comparative study of the effectiveness of ensemble technique for sentiment classification, namely the part-of-speech based feature sets and the word-relation based feature sets. Then, three text classification algorithms, namely naive Bayes, maximum entropy and support vector machines, are employed as base-classifiers for each of the feature sets to predict classification scores.

Long-Sheng Chen (2011) proposed a neural network based approach, which combines the advantages of the machine learning techniques and the information retrieval techniques.

Naive Bayes is a simple but effective classification algorithm. The Naive Bayes algorithm is widely used algorithm for document classification (Melville et al., 2009; Rui Xia, 2011; Ziqiong, 2011; Songho tan, 2008 and Qiang Ye, 2009). The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories.

In most of the comparative studies it is found that SVM outperforms other machine learning methods in sentiment classification. Ziqiong Zhang (2011) showed a contradiction in the performance of SVM. They focused their interest on proving that the chosen machine learning model could be able to draw its own conclusion from the distribution. Despite its unrealistic independence assumption, the naive Bayes classifier surprisingly achieves better performance than SVM.

Zhu Jian (2010) proposed an individual model based on Artificial neural networks to divide the movie review corpus into positive , negative and fuzzy tone which is based on the advanced recursive least squares back propagation training algorithm.

**2.1 Domain Explanation**

**Artificial Intelligence:**

Artificial intelligence (AI) is the intelligence exhibited by machines or software. It is also the name of the academic field of study which studies how to create computers and computer software that are capable of intelligent behavior. Major AI researchers and textbooks define this field as "the study and design of intelligent agents", in which an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. John McCarthy, who coined the term in 1955, defines it as "the science and engineering of making intelligent machines". AI research is highly technical and specialized, and is deeply divided into subfields that often fail to communicate with each other. Some of the division is due to social and cultural factors: subfields have grown up around particular institutions and the work of individual researchers. AI research is also divided by several technical issues. Some subfields focus on the solution of specific problems. Others focus on one of several possible approaches or on the use of a particular tool or towards the accomplishment of particular applications. The central problems (or goals) of AI research include reasoning, knowledge, planning, learning, natural language processing (communication), perception and the ability to move and manipulate objects. General intelligence is still among the field's long-term goals. Currently popular approaches include statistical methods, computational intelligence and traditional symbolic AI. There are a large number of tools used in AI, including versions of search and mathematical optimization, logic, methods based on probability and economics, and many others. The AI field is interdisciplinary, in which a number of sciences and professions converge, including computer science, mathematics, psychology, linguistics, philosophy and neuroscience, as well as other specialized fields such as artificial psychology.

The field was founded on the claim that a central property of humans, human intelligence—the sapience of Homo sapiens—"can be so precisely described that a machine can be made to simulate it." This raises philosophical issues about the nature of the mind and the ethics of creating artificial beings endowed with human-like intelligence, issues which have been addressed by myth, fiction and philosophy since antiquity. Artificial intelligence has been the subject of tremendous optimism but has also suffered stunning setbacks. Today it has become an essential part of the technology industry, providing the heavy lifting for many of the most challenging problems in computer science.

**Machine Learning:**

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions. Machine learning is closely related to computational statistics; a discipline that aims at the design of algorithm for implementing statistical methods on computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR), search engines and computer vision. Machine learning is sometimes conflated with data mining, although that focuses more on exploratory data analysis. Machine learning and pattern recognition "can be viewed as two facets of the same field." When employed in industrial contexts, machine learning methods may be referred to as predictive analytics or predictive modelling.

Types of problems and tasks

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are:

* Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
* Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
* Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

Between supervised and unsupervised learning is semi-supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing. Transduction is a special case of this principle where the entire set of problem instances is known at learning time, except that part of the targets are missing.A support vector machine is a classifier that divides its input space into two regions, separated by a linear boundary. Here, it has learned to distinguish black and white circles. Among other categories of machine learning problems, learning to learn learns its own inductive bias based on previous experience. Developmental learning, elaborated for robot learning, generates its own sequences (also called curriculum) of learning situations to cumulatively acquire repertoires of novel skills through autonomous self-exploration and social interaction with human teachers, and using guidance mechanisms such as active learning, maturation, motor synergies, and imitation. Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system:

In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam". In regression, also a supervised problem, the outputs are continuous rather than discrete. In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task. Density estimation finds the distribution of inputs in some space. Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics.

**2.2 Existing Solution**

Naïve Bayes Classification

* D : Set of tuples
* Each Tuple is an ‘n’ dimensional attribute vector
* X : (x1,x2,x3,.... xn)
* Let there be ‘m’ Classes : C1,C2,C3...Cm
* Naïve Bayes classifier predicts X belongs to Class
* Ci iff
* P (Ci/X) > P(Cj/X) for 1<= j <= m , j <>i
* With many attributes, it is computationally expensive to evaluate P(X/Ci).
* Naïve Assumption of “class conditional independence” is P(X/Ci) = P(x1/Ci) \* P(x2/Ci) \*...\* P(xn/ Ci)

**2.3. H/W & S/W Requirements**

For implementation of our application we need various software and hardware configuration.

* + - 1. **Hardware Requirement:**

1. CPU: 34 bit / 64-bit
2. Processor: Intel i5 (4th Generation)
3. RAM: 8 GB or more
4. Hard Disk: 500 GB or more
   * + 1. **Software Requirement:**
5. Operating System: Windows 7
6. Software: Python 2.7
7. Natural Language Toolkit (NLTK)
8. Twitter REST API
9. OAuth 2.0
10. Google Charts