VSM Based Classification of Data Objects with Individual Treatment of Continuous and Discrete Attributes

KomalKumar Bhatia
Associate Professor, CSE Deptt.
YMCAUST, Faridabad, Haryana, India

AtulSrivastava
M. Tech. (IT)
YMCAUST, Faridabad, Haryana, India

VeenaGarg
M. Tech. (CE)
YMCAUST, Faridabad, Haryana, India

ABSTRACT
Classification is a technique, used in data mining, for identification of membership of a particular data object. In this paper we provide a technique of classification that is an enhancement of an existing method of information retrieval i.e. Vector Space Model. Vector space model is applied on text data and generally used to determine the relevance of query to the web pages in information retrieval. Data objects are categorized in two communities based on their attributes, one having discrete-valued attributes and second having continuous-valued attributes. In almost every previous attempt in this area has treated both of the communities of data objects separately. For scalability point of view of the classifier one type (discrete/continuous) is converted to the other (continuous/discrete).This conversion sometimes may hamper the accuracy. But in this paper continuous and discrete attributes are treated individually without tempering their representation. This paper emulates VSM to be used for classification in the same way it is used for determining query relevance in information retrieval. The results show that the enhanced model achieved very good results in performance and the setup time is also satisfactory for a large collection of data objects.
This paper is organized as section 1 contains the basic terminology about classification and introduction of vector space model, section 2 contains the related work that has already been done in literature, section 3 contains model construction for classification i.e. simulation of existing vector space model for information retrieval and use of this model for classification of unseen data tuple, section 4 contains pseudo code for VSM classification. Section 5 shows experiment and results analysis through an example. Section 6 concludes the paper and throws light on future aspects.

Keywords:
Information retrieval, Vector space Model, Classification, Continuous attributes, Discrete attributes, Classification technique.

1. INTRODUCTION

1.1 Classification

Classification is a form of data analysis that can be used to extract model describing important data classes or to predict future data trends5. Classification predicts categorical labels. Many classification methods6 have been developed in machine learning. Some of them include k-nearest neighbor classifier, Decision tree classifier, Bayesian belief network, and classification by back propagation etc. Data classification is a two-step process; first step is learning step, where classification algorithm builds classifier by analyzing training dataset made up of database tuples and their associated class labels, and second step determines the class label for unseen data tuple usingthis classifier.

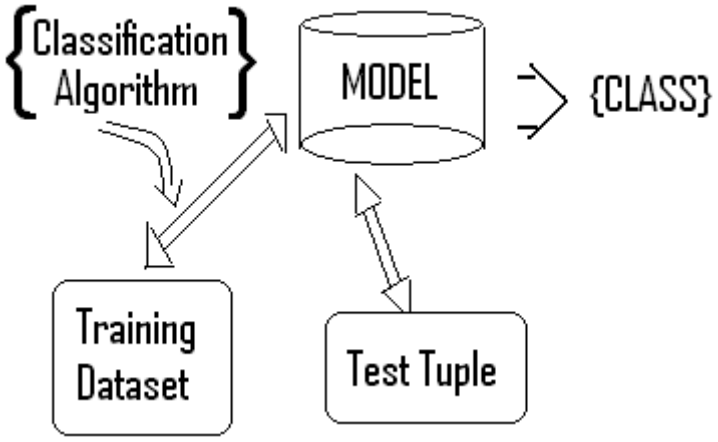

Fig 1: Supervised classification Architecture

1.2 Vector Space Model

1.2.1 Vector Space Model construction

The Vector Space Model (VSM) is a standard technique in Information Retrieval in which documents are represented through the words that they contain. It was developed by Gerard Salton in the early 1960's to avoid some of the information retrieval problems. Vector spaces models convert texts into
matrices and vectors, and then employ matrix analysis techniques to find the relations and key features in the document collection. It represents queries and documents as vectors of terms which can be words from the document or the query itself. The most important thing is to represent relevance between documents in this information space, which is achieved by finding the distance between the query and the document1.
The weight of relevance of a query in the document can be calculated using some similarity measures such as cosine or dot product or other measurement.
The VSM relies on three sets of calculations. This model can work on selected index of words or on full text. The calculations needed for the vector space model are:

1. The weight of each indexed word across the entire document set needs to be calculated. This answers the question of how important the word is in the entire collection.

2. The weight of every index word within a given document (in the context of that document only) needs to be calculated for all N documents. This answers the question of how important the word is within a single document.

3. For any query, the query vector is compared to every one of the document vectors. The results can be ranked. This answers the question of which document comes closest to the query, and ranks the others as to the closeness of the fit. The weight can be calculated using this equation3:

$$w_i = tf_i * log\left(\frac{D}{df_i}\right) \quad \text{..... (1.2.1)}$$

Where,
• tfi = term frequency (term counts) or number of times a term i occurs in a document. This accounts for local information.
• dfi = document frequency or number of documents containing term i
• D = number of documents in a database. The D/dfi ratio is the probability of selecting a document containing a queried term from a collection of documents. This can be viewed as a global probability over the entire collection. Thus, the log (D/dfi) term is the inverse document frequency, IDFi and accounts for global information.

1.2.2 Similarity measurement

There are many different methods to measure4 how similar two documents are, or how similar a document is to a query in VSM. These methods include the: cosine, dotproduct, Jaccard coefficient and Euclidean distance. Here cosine measure has been used which is the most common.

Suppose Q is the query and $\mathbf{D}_i$ is the document, where $1 \leq i \leq n$, vectors for Q and $\mathbf{D}_i$ are calculated using above formula for weight calculation (1.1). Similarity of query to each document is calculated by using the following formula3,

$$\mathrm{Sim}(Q,D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2}\sqrt{\sum_i w_{i,j}^2}} \quad \text{…… (1.2.2)}$$

2. RELATED WORK

Data mining involves many sophisticated tools for analysis of unknown valid patterns and their relationship in the large dataset. Classification is capable of processing a wide variety of data than other regression methods and therefore it is growing in popularity.
Many attempts by researchers have been made to rectify the problem of classification and they have come up with excellent results in terms of accuracy and scalability. One approach is Decision trees6. The instances are classified by sorting them based on their feature values. An instance to be classified is represented by the node in the decision tree as feature. Each branch represents a value that the node can assume. Classification of instances starts at the root node and is further sorted based on their feature values. Decision trees usually do not show variations since they use sorting based on a single feature value at each internal node. But sometimes decision tree algorithms cannot perform well with problems when diagonal partitioning is required in the dataset.
Another technique is Bayesian Network6 (BN) that is a graphical model that reflects probability relationships among a set of variables features of the data instances. The Bayesian network structure is a directed acyclic graph (DAG) in which nodes represent different features and there is one-to-one correspondence between them. The edges in DAG represent influence of one feature on another whereas no edge implies conditional independence. A feature is also not dependent on its non-descendent node (feature). Bayesian Network is not suitable for the dataset having large number of features (Cheng et al., 2002). Because constructing a very large network incurs huge cost in terms of time and space. Another problem with BN is that all the numerical features need to be discredited in most cases.
K-Nearest neighbor classifiers6 are based on analogical learning. The data objects are treated as training samples with n-dimensional numeric attributes. Each sample is represented in n-dimensional space as a point. All the known samples are stored in the n-dimensional pattern space in such manner and are called training samples. When an unknown sample is to be classified it is also represented in the same n-dimensional pattern space and a k-nearest neighbor classifier then searches for k training samples which are closest to the given unknown sample. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

Much more research has been done that is based on the above techniques like Rule based classification technique is based on Decision Tree method, variants of Bayesian network are Bayesian belief network. Similarly Instance based learning is based on K-nearest neighbor technique.

Almost every attempt so far has classified data objects with same type of attributes (continuous or discrete) by converting one type to another. This conversion discredits the actual representation of attribute value and sometimes may cause unpredictable inaccurate results. This paper mainly focuses on this problem of conversion that compromises with accuracy and provides a method in which continuous and discrete attributes are treated individually without converting them into same type.

3. VSM BASED CLASSIFIER

Supervised classification techniques involve a two step process; first model construction based on identified data objects, also called learning phase, and then identification of class label of unknown data object using this model. This paper also presents a method that follows supervised learning phenomenon.

3.1Model construction

In Vector Space model term frequency and document frequency are used for weight calculation. But term frequency and document frequency used in the formula (Eqn. 1.2.1) for weight calculation cannot be used here. Because term frequency is the number of times term appeared in a document and document frequency is measurement of term's presence in complete data set. But in our case classification is done for the dataset with tuples having continuous and discrete attributes. In such case concept of frequency needs to be modified. Value of continuous attribute and local probability of discrete attribute can be considered as term frequency as well as document frequency for inverted document frequency.

Therefore weight vector i^{th} tuple X_i can be given as:

$$v_i = [w_{i,1}, w_{i,2}, \ldots\ldots\ldots\ldots, w_{i,m}]^T$$

Here for Continuous attributes

The formula for $W_{i,j}$ is hence modified as follows:

$$W_{i,j} = weight_{i,j} * \log\left(\frac{max_j}{weight_{i,j}}\right) \quad \ldots\ldots (3.1)$$

Where,

$weight_{i,j}$ is continuous value of j^{th} attribute of i^{th} tuple.

max_j is maximum permissible value for j^{th} attribute in the dataset.

For Discrete attributes the formula for $W_{i,j}$ is hence modified as follows:

$$W_{i,j} = Lp_{i,j} * \log\left(\frac{Gp_{i,j} + Lp_{i,j}}{Lp_{i,j}}\right) \quad \ldots\ldots (3.2)$$

Where,

$Lp_{i,j}$ is local probability that a tuple has discrete value $V_{i,j}$ for j^{th} attribute, i.e.

$Lp_{i,j} = \frac{|V_{i,j}|}{|D|}$, { $|V_{i,i}|$ is number of tuples having discrete value $V_{i,j}$ for j^{th} attribute, and $|D|$ is total number of tuples in training dataset.}

$Gp_{i,j}$ is global probability that a tuple has discrete value $V_{i,j}$ for j^{th} attribute. It is experimentally measured.

3.2 Similarity measurement

After model construction the next step is to classify the unseen data tuples using the model. To determine the class of unseen data tuple same similarity measure is used as used in Vector Space Model in information retrieval.

Now, suppose X^T is the unseen data tuple (test tuple) to be classified and X_i is the training set tuple, where $1 \leq i \leq n$, vectors for X^T and X_i are calculated using above formulae for weight calculation (Eqn. 2.1). Similarity of test tuple to each training tuple is calculated by using the following formula,

$$Sim(X^T, X_i) = \frac{\sum_i w_{X^T,j} w_{i,j}}{\sqrt{\sum_j w^2_{X^T,j}} \sqrt{\sum_i w^2_{i,j}}}$$

……..(3.3)

This similarity is then used to determine the class of the test tuple. The probability that X^T will belong to the same class to which X_i belongs is proportional to the value of $Sim(X^T, D_i)$, i.e. higher the value of $Sim(X^T, D_i)$, closer X^T and X_i will be.

4. VSM CLASSIFICATION ALGORITHM

```
D -Training Data Set
Xi -ith Data tuple in the training data set
Yf -Class Associated with with ithtuple of training data set
XT -Unseen Test Tuple
1   begin
2     for i ← 1 to n          //n is total number of tuples in training dataset
3       for j ←1 to m         //m is number of attributes in each tuple of the training dataset
4         if jth attribute is discrete
5           W(i,j) = Lp(i,j) * log((Gp(i,j) + Lp(i,j)) / Lp(i,j))
6         else if jth attribute is continuous
7           W(i,j) = weight(i,j) * log(max_j / weight(i,j))
8       end
9       vi = [w(i,1), w(i,2), ....., w(i,m)]^T
10    end
11    for test tuple XT
12          calculate v^T = [w(XT,1), w(XT,2), ....., w(XT,m)]^T
13    for i←1 to n
14          sim[i] = Sim(XT, Xi) = Σ_i w(XT,j) w(i,j) / (sqrt(Σ_j w^2(XT,j)) sqrt(Σ_i w^2(i,j)))
15    end
16    max = -9999;
17    for i←1 to n
18          if(max<sim[i])
19                max = sim[i]
20                class(XT) = Yf
21    end
22    return class(XT)
23  end
```

The Algorithm works incrementally; initially there is training dataset D having 'n' attributes and 'm' tuples.

The main steps of the algorithm are:

Line [1 – 8]: This section calculates weights corresponding to each attribute for each tuple. Line 4 and 6 decide the type of attribute based on which Line 5 and 7 calculate weights.

Line [9]: It defines the weight vector corresponding to i^{th} tuple.

Line [11-12]: Calculates weight vector for unseen test tuple.

Line [13-14]: Calculates similarity of test tuple to each training tuple.
Line [17-23]: In Line 17-21 analyse the similarity calculated above and decides the closest training tuple to the test tuple and Line 22 returns final class label for test tuple.

In the above algorithm, line 1 to 10 incur time complexity of O(n2), line 11, 12, 16 & 22 have constant time complexity, and line 13 to 15 and 17 to 21 take O(n) time. Therefore the overall time complexity of the algorithm is O(n2).Overall space complexity of the algorithm is also O(n2).

5. EXPERIMENT AND RESULT ANALYSIS

The proposed method is rigorously tested against the dataset "Bank Loan Officer Analysis", which contains continuous attribute 'Age' and discrete attribute 'Income' (LOW/ MEDIUM/ HIGH). On the basis of attribute values the officer has to take decision about the customer that whether the customer is SAFE or RISKY. We implemented the proposed model in 'C'. First training dataset with defined class labels is used to construct model and then it is used to classify unseen data tuples i.e. new customers. The working of proposed model has been depicted by an example as follows:

Table 1: Training Data Set

SNo	Name	Age (Years)	Income	Loan Decision
1	Sandy	25	LOW	RISKY
2	Lee	35	MEDIUM	RISKY
3	Rick	42	HIGH	SAFE
4	Susan	57	LOW	SAFE
5	Joe	23	HIGH	RISKY
6	Claire	67	MEDIUM	SAFE
7	Smith	41	LOW	RISKY
8	Phips	70	LOW	SAFE
9	Yuki	76	HIGH	SAFE
10	Juan	48	LOW	RISKY
11	Sylvia	33	HIGH	RISKY
12	Anne	27	MEDIUM	RISKY
13	Bello	82	HIGH	SAFE
14	Henry	39	MEDIUM	SAFE

Step 1:Model Construction
In this step vectors for each data tuple of training data set is calculated using Eqn. 3.1 and 3.2.
Max value for age is experimentally taken 120 years.

Table 2: Weight vector matrix

$W_{I,J}$ is j^{th} weight in vector for i^{th} tuple		
S. No. (i)	$W_{i,1}$	$W_{i,2}$
1	0.1703	0.1124
2	0.1873	0.1174
3	0.1915	0.0604
4	0.1843	0.1124
5	0.1650	0.0604
6	0.1696	0.1174
7	0.1912	0.1124
8	0.1639	0.1124
9	0.1508	0.0604
10	0.1910	0.1124
11	0.1850	0.0604
12	0.1749	0.1174
13	0.1356	0.0604
14	0.1904	0.1174

Step 2: Classification of test tuple

Suppose test tuple X^T <age= 37, income= LOW> is to be classified. Vector for test tuple is calculated by using same formula in Eqn. 3.1 for continuous attributes and Eqn. 3.2 for discrete attributes.

$$V^T = [0.1891, 0.1124]$$

Table 3: Similarity values

$Sim(X^T, X_i)$ is similarity of X^T to each tuple X_i in the dataset	
S. No. (i)	$Sim(X^T, X_i)$
1	0.99893
2	0.99963
3	0.97353
4	0.99878
5	0.98297
6	0.99771
7	1.0000
8	0.99799
9	0.98807
10	1.0000
11	0.97584
12	0.99859
13	0.99323
14	0.99997

Step 3:Results and Evaluation

Table 3 shows that $Sim(X^T, X_7)$ has highest value so X^T will belong to the same class to which X_7 belongs. X_7 i.e. class of unseen data tuple X^T will be "RISKY".

Experiments were run on a number of unseen data tuples to classify in the data set. After observing the data set retrieved results are correct.
This algorithm has been tested on different data set having continuous and discrete attributes and it is found that the algorithm successfully classified all the unseen data tuples.

6. CONCLUSION AND FUTURE WORKS

In this paper a modified model of existing VSM for information retrieval has been introduced, which is applied for the classification of continuous and discrete objects. The modified method achieved good result for classifying unseen data tuples. The results yielded by the experiments prove that this approach seems very competitive.
There are several research directions that look attractive for future exploration. Some of them are: to apply modified method of VSM to experience with more similarities measures. We also intend to explore performance measure technique for comparison with other families of classification methods.

7. REFERENCES

[1] Van Rijsbergen, Keith,"Information Retrieval", Butterworths London, 1979.

[2] M.J. Xavier, Sundaramurthy, P.K. Viswanathan, G. Balasubramanian ,"Improving prediction accuracy of loan default- A case in rural credit".

[3] "Vector space model –Wikipedia ", http://en.wikipedia.org/wiki/Vector_space_model

[4] "Scoring, Term Weighting and the Vector Space Model", www.stanford.edu/class/cs276/handouts/lecture6-tfidf.ppt.

[5] "Statistical classification (machine learning)", http://en.wikipedia.org/wiki/Classification_(machine_learning).

[6]Thair Nu Phyu , Survey of Classification Techniques in Data Mining,2009.

[7]AtulSrivastava,VeenaGarg,"An Adaptation of Vector Space Model for Classification of Continuous data objects", 2011.