

Social Learning under Uncertainty Based on Dempster-Shafer Approach for Minimizing True Error of Machine Learning

Abstract

Minimizing true error of the classification process under uncertainty is one of the difficult issues in the field of machine learning. Researchers do not address this topic until this time despite its importance in practical life. This paper can be considered as a development of the concept of social learning presented the intellectual leap in the machine learning area as given before by the authors. Novelty in this paper is to present a new approach that can deal with the conditions of uncertainty resulting from multiple sources. This paper also presents a new method of social learning based on benefits offered by the Dempster-Shafer theory (DST) of evidence. The paper provides experimental results on six benchmarks. The results attained from the comparison using six benchmarking problems illustrate a superior performance of the proposed method compared with the best results attained in the literature of machine learning domain till now.

Keywords: Uncertainty, Social Machine Learning, Dempster-Shafer Theory, True Risk, Vapnik – Chervonenkis Theory.

1. Introduction

Russian school provided a major development in the field of machine learning. This school is based on a deep and strong background of mathematics. This ideology be reflected through the adage of Vapnik" nothing more practical than a good theory." Vapnik (1964) gives a leading Russian school in the field of machine learning. He has opened new horizons by his work depending on the structure of the machine learning, not the data itself as it was a strong belief in past. Pioneers of Russian school in this field are Vapnik and Chervonenkis who introduced a new theory called Vapnik and Chervonenkis theory (VC-dimension). VC theory has three dimensions: conceptual, mathematical, and constructive learning. Firstly, the conceptual dimension is presented and developed only by Vapnik (1964). This dimension is concerned with basic characteristics of inference from finite samples based on the idea of Empirical Risk Minimization (ERM). Secondly; the mathematical dimension is presented and demonstrated by Vapnik and chervonenkis(1974).

They proved the following inequality that describes an upper bound of classification error based on the structure of machine learning for unseen data.

$$\mathcal{E}(f) \leq \mathcal{E}_n(f) + \sqrt{\frac{h \left(\log\left(\frac{2n}{h}\right) + 1 \right) - \log\left(\frac{\eta}{4}\right)}{n}} \quad \dots (1) \quad (\text{Vapnik and Chervonenkis, 1974}).$$

Where $\mathcal{E}(f)$, $\mathcal{E}_n(f)$ denote the true error and empirical error. The probability that this bound is crashed equals η , machine learning capacity (VC dimension) of classifiers is h . VC confidence is the second term (the square root) of the inequality (1) (Vapnik and Chervonenkis, 1974). It is worth mentioning that these two dimensions be well known in the seventies of the last century. The main dilemma of the mathematical dimension is the inability to find a method clear and specific to how to measure VC-dimension for various machine learning. Inequality (1) reflected the general shape but failed to explain how to find

VC-dimension to each machine learning. So, it could not answer many questions about practical challenges in applying the theory. This deficit has weakened many of the importance of this theory in practice and application.

Thirdly, the constructive dimension which is extended to the mathematical dimension. This dimension is concerned with taking a suitable complexity model to fit data to attain an acceptable bound of the error through determining VC-dimension for different types of machine learning. One of the Pioneers in the constructive dimension is Girosi (1995) who presents a new technique for estimating VC dimension bounds based on the theory of statistical learning. In the same direction Kon and Raphael (2005) enhance the paper of Girosi (1995) for finding the L1-norm of upper bound function. Later, Kon and Raphael (2006) develop Girosi's approach (1995) for existing approximate bound of the error based on Hilbert spaces. This method is called Reproducing Kernel Hilbert Spaces. Key et al. (1999) who pay their attention for determining the bound of the error using Bayesian decision-theoretic trend. Also, Miller (2002) presents a new mechanism for selecting the variables in the model of regression that minimizes the approximate bound of the error. In the same direction, Teytaud and Lallich(2001) display the mechanism of using an appropriate VC-dimension to be a restricted bound on the accepted risk from database. This method is named "association rules." This area attracts many researchers who contribute various methods for estimating the upper bound based on the theory of statistical learning. They enhanced the accuracy of the bound by presenting a new concept called Rademacher's complexity but this method allows the curse of dimensionality (Mendelson 2003). Vapnik and Chervonenkis (1998) present the new concept of Structural Risk Minimization (SRM) which displays a trade-off between the complexity of the bias function and the accuracy of the approximation and. Also, Onshuus, and Usvyatsov (2011) propose a method to give the uniform bound on VC. In the same trend, the uniform bound on VC technique has more attention through the work of Shelah et al. (2008).

One of the recent applications is presented by Matteo Riondato and Fabio (2013) for estimating the sample complexity bound which is based on empirical VC-Dimension to the field of data mining. The constructive dimension presents a variety work for estimating VC-dimension for different types of machine learning as Support vector machines, Decision tree, gene expression programming.....Etc. However, the previous constructive work has the following Weaknesses:

- 1-Absence of the work that is based on social learning.
- 2- Majority of the work is devoted to developing a neural network area.
- 3- Assuming all machine learning problems under certainty conditions away from realism, but there is no academic work explains how to deal under conditions of uncertainty.
- 4- Determining the optimal structural based on a criterion (accuracy level) .This trend can be considered as monomeric outlook does not take into account other dimensions such as CPU time.
- 5- Also it cannot perform minimizing both terms of Structural Risk Minimization (SRM) simultaneously.

These weaknesses are considered the main motivator for many researchers to tackle these challenges. Recently, Zaher et al.(2013) introduced a new trend called social machine learning. They tackle the first three challenges through proposing new approach is called Tropical Collective Machine Learning (TCML). This approach is the first academic research introduces the concept of social machine learning in this field.

This paper presents a novel approach for tackling the social learning under uncertainty conditions. This approach is based on Dempster-Shafer theory for dealing with subjective probability by applying on six machine learning techniques. The main merit of the proposed approach is the ability to attain an empirically and an approximately true error by determining

suitable model complexity under uncertainty. The following sections of this paper is structured as follow; Section two present a brief overview of the Dempster-Shafer used as a base for the experimental work. Section three displays the proposed approach and gives the used algorithm. Section four shows the data sets used simulation Procedure; performance measures, and display experiments and results. Section five displays the discussion and Section six conclusion of the paper.

2. Dempster-Shafer Evidence Theories

The theory is considered as development and generalizations of the Bayesian theory provided by Dempster (1967) then the improvements are added to this theory by Shafer (1976). The core of this theory is the evidence theory which based on the fundamental concept named Basic Belief Assignment (BBA) that has a clear weakness that is inability to introduce a model of probability in a classical way; (Joseph, and Gary, 2005). This theory is composed of three fundamental functions. The first function is named the basic belief assignment (BBA). The second function is named Belief Function (Bel). Finally, the third function is named Plausibility (Pl). The BBA, denoted by $m(\emptyset)$, determine a mapping function convert the power set into the interval between zero and one. The BBA is presented by the equations:

$$m(X) \rightarrow [0,1] \quad (1)$$

$$m(\emptyset) = 0 \quad (2)$$

$$\sum_{A \in X} m(A) = 1 \quad (3)$$

(Rombaut, and Zhu, 2002)

The merit of this theory is presenting the concept of the ignorance, also assigning certain value for it. In real life, data fusion is required where multiple sources of information are combined to inform the best judgment of the situation. The main gain from information accumulation is the ability to simply and summarizes the collected data to gain benefits of multiple sources of information. There are many types of link sources of information, which is called "Combination rules.". The real benefit is generated from multiple sources are presenting different measures for the same frame of discernment. The fundamental assumption of the Dempster -Shafer theory that sources are used to be independent. Dempster introduced his combination rule (for two sources) as follows:

$$m^{1,2}(C) = \frac{\sum_{A \cap B = C} m^1(A) m^2(B)}{\sum_{A \cap B \neq \emptyset} m^1(A) m^2(B)} = \frac{\sum_{A \cap B = C} m^1(A) m^2(B)}{1 - \sum_{A \cap B \neq \emptyset} m^1(A) m^2(B)} \quad (\text{Tazid Ali et al., 2012})$$

3. The Proposed Approach Dempster-Shafer Collective Machine Learning (DSCML)

The proposed approach (DSCML) is composed of two consecutive stages. The first stage is to apply the technique called Tropical Collective Machine Learning (TCML) to select the best

appropriate structure of machine learning that will be used in the next stage (Zaher et al., 2013). The second stage is based on Dempster-Shafer approach for dealing with uncertainty of subjective probability of collective machine learning to attain minimum true error. TCML technique is an intelligent computational method that calculates the expected structure cost of machine learning. TCML algorithm, given by Zaher et al. (2013), is illustrated as follows:

1- Run the machine learning to classify the data.
 2- Determine the relative importance of variables at input n -dimensional.
 3- Build $A = [a_{ij}]$ such $1 \leq i \leq n, 1 \leq j \leq n$ the matrix.
 Whose entries are the cost structure of VC dimension to arc_{ij} which means sequential hybridization from machine learning I to machine learning j .
 4- Build $B = [b_{ij}]$ such $1 \leq i \leq n, 1 \leq j \leq n$ the matrix whose entries are empirical error obtained by sequential hybridization from machine learning I to machine learning j .
 5- Build $C = [c_{ij}] = A \otimes b$ such $1 \leq i \leq n, 1 \leq j \leq n$ the matrix
 Whose entries are true error obtained at each arc_{ij} .
 6- Build $C^1, C^2, C^3, \dots, C^{n-1}$
 Which $C^1 = C, C^2 = C^1 \otimes C^1, C^3 = C^1 \otimes C^2, C^4 = C^1 \otimes C^3, C^5 = C^1 \otimes C^4$.
 7- Find $C^* = C^1 + C^2 + C^3 + \dots + C^{n-1}$
 8- choose the minimum value in matrix $C^* [c_{ij}] = C_{ij}$ which is called C^*
 Such that $1 \leq i \leq n, 1 \leq j \leq n$.
 9- go to the next iteration, find a_{ij} to arc_{ij} .
 9- Calculate upper bound $u = C^* - a_{ij}$ for each arc_{ij} .
 10- If $b_{ij} > C^*$ for $1 \leq i \leq n, 1 \leq j \leq n$ stop, Else $n=n-1$ go to step 1
 11- Find the total error $= b_{ij}$

The first stage: TCML algorithm, Zaher et al. (2013)

1- Run the machine learning model to classify data.
 2- Determine the weight of the relative importance for every variable.
 3- construct the weight matrix $A = [a_{ij}]$ such
 $1 \leq i \leq n, 1 \leq j \leq n$ the matrix. This matrix presents the relative importance of each variable for the first source (first machine learning model).
 4- Construct the matrix $b = [b_{ij}]$ such $1 \leq i \leq n, 1 \leq j \leq n$ the matrix. This matrix presents the relative importance of each variable for the first source (first machine learning model).

5- Construct the matrix C . This matrix is formulated by combined belief function achieved by the two belief function.

$$m(Z) = \frac{\sum_{A \cap B = Z \neq \emptyset} m(A) \cdot m(B)}{1 - \sum_{A \cap B = \emptyset} m(A) \cdot m(B)},$$

5-estimate the final true error achieved at this dimension.

6-if the true error satisfies the stopping criteria, move to step 8, else move to step 7.

7-Eliminate the dimension with the lowest relative weight and replace n by $n-1$ then Go to step 1.

8- Output the final structure of the model.

The second stage of proposed algorithm DSCML

3.1 Data Sets (benchmarking problems)

This subsection displays six types of benchmark of classification that be used in the experiments.

Table 1.The data of benchmark of classification that be used in the experiments.

<i>Data set</i>	<i>Bench mark name</i>	<i>No. of Variables</i>	<i>No. of rows</i>
<i>first</i>	Vowel	10	990
<i>second</i>	Telugu Vowel	3	871
<i>third</i>	Wisconsin Breast Cancer	9	699
<i>fourth</i>	Heptitates	19	155
<i>fifth</i>	Cleveland heart Disease	13	303
<i>Sixth</i>	Diabetes	8	766

These datasets used in the experiments are generated from UCI machine learning database.

<http://www.ics.uci.edu/~mllearn/MLRepository.html> (20 July 2013)

3.2 Performance Measures

Accuracy level is used to measure the accuracy of the proposed approach.

CPU time is a second measure to indicate the consumed time by the algorithm.

3.3 Simulation Procedure

The simulation procedure is used in the experiments based on dividing the whole data set, fifty percentage of them used for training and thirty percentages of them used for validating, twenty percentages of them used for testing. These experiments produced: 20 times repeated; the foldcross validation is 10.

4. The Experiments &Results

This section is divided into two subsections. The first of them is assigned for computing the VC – dimension of various techniques of machine learning. The second subsection is assigned for finding the optimal number of dimensions chosen for minimizing the true risk.

4.1 Computing the VC-Dimension

Based on the general inequality (1), the second term of the inequality called (structure cost of machine learning) depends on three variables h ; is VC- machine learning complexity, n is the sample size and η : is the accuracy level required by the machine learning. It is logically through the process of minimization the controlled variable is h . It is worth mentioning the existence of a positive relationship between the data number of dimensions and VC-dimension. The appropriate structure for various machine learning techniques using the proposed approach with confidence level 0.9999 is shown in the following table.

Table 2. The Optimal structure for machine learning techniques used in the experiments

Technique of Machine Learning	Formula of calculating VC-dimensions	Controlling Variables for appropriate structure by DS-ML	VC dimension	Calculated Optimal Structure Cost
Single Decision Tree (STR)	$[\log_2(n - d + 2)] + d$. (Olcay Taner Yilidiz, 2012)	$n=20, d=3$	5.726833	0.062079
Multilayer Perceptron (MLP)	$VC = 2\lceil \frac{K}{2} \rceil n$ (Eric B.Baum. et al, 1988)	$K=21$ $n=3$	63	0.026055
Radial Basis Function (RBF)	$N \log(h)$ (Sakurai A, 1993)	$N=18$ $h=4$	43.34832	0.028685
Linear classification (Linear)	$VC = n + 1$ (Vapnik, 1999)	$n=16$	17	0.039216
Support vector machine (SVM)	$h \leq \min(R^2 A^2, n) + 1$ (Marcin Owczarczuk, 2009)	$R=0.01667, A=2000$ $n=18$	19	0.037604
Group Method of Data Handling (GMDH)	C_h^{n+h-1} (Sakurai A, 1993)	$n=12$ $h=3$	55	0.026927

4.2 A Comparison between the Best Results Achieved in the Literature of Machine Learning Field and TCML.

This subsection presents a comparison between the TCML technique and the best results achieved in the literature of machine learning field.

<http://duch-links.wikispaces.com/Classification+results>

The type of data set	The best Method achieved best results for data set	Accuracy level achieved by the Reference	accuracy level of TCML	CPU time estimated in seconds for TCML
Vowel	CART-DB, 10xCV on total set !!!	90.0 Shang, Breiman	93.1	24.85
Telugu Vowel	3-NN, Manhattan	87.8 Kosice	90.8	29.18
Wisconsin Breast Cancer	FSM	98.3 (RA)	99.18	37.98
Heptitates	CMLP2LN/SSV single rule	76.2 WD/K. Grabczewski stable rule	78.2	31.65
Cleveland heart Disease	IncNet+ transformations	90.0 Norbert Jankowski; check again!	92.8	32.38
Diabetes	Logistic discrimination	77.7 Statlog	78.8	16.98

Table 3. The results achieved by the TCML compared with the best results obtained in the literature of machine learning field.

Table 3 shows the superior performance of TCML compared by other techniques this dominant performance due to the strength points of the technique:

- 1- Applying the concept of social learning that maximizes benefits generated from the integration of the information.
- 2- Specifying the suitable model complexity accurately which leads to minimizing the first term of Vapnik formula.
- 3- The technique TCML can simultaneously minimize both terms of Vapnik formula. (Zaher et al., 2013)

4.3A Comparison between DSCML and TCML

Subsection presents a comparison between the TCML technique and DSCML supported by T-TEST for the accuracy and the CPU time.

Table 4. The results achieved by the DSCML compared with TCML according to classification accuracy.

The type of data set	accuracylevel of TCML	accuracylevel of DSCML	Improvement percentage	P –value
Vowel	93.1	94.39	1.29%	0.001**
Telugu Vowel	90.8	91.27	0.47%	0.006**
Wisconsin Breast Cancer	99.18	99.47	0.29%	0.008**
Heptitates	78.2	79.9	1.7%	0.000**
Cleveland heart Disease	92.8	93.4	0.6%	0.005**
Diabetes	78.8	80.8	2%	0.000**

Table 4 shows the superior performance of DSCML compared by TCML according to the accuracy level that proved by T-TEST at the significance level is 1%.

Table 5. The results achieved by the DSCML compared by TCML in classification accuracy.

The type of data set	TCML CPU time estimated in seconds	DSCML CPU time estimated in seconds	Improvement percentage	P –value
Vowel	49.12	24.85	49.5%	0.000**
Telugu Vowel	39.76	29.18	26.67%	0.000**
Wisconsin Breast Cancer	55.19	37.98	31.2%	0.000**
Heptitates	46.34	31.65	32%	0.000**
Cleveland heart Disease	33.19	32.38	2.5%	0.009**
Diabetes	27.80	16.98	39%	0.000**

Table 5 shows the superior performance of DSCML CPU time compared by TCML according to CPU time that proved by T-TEST at the significance level is 1%.

5. Discussion

The paper has demonstrated the following results:

- There are potential gains available through using of the DSCML approach.
- Found statistically superior performance related to the allowance for ignorance/uncertainty in the judgments.

-The proposed approach can merge multiple resources simultaneously.

6. Conclusion

Results that have been obtained proved beyond any doubt the ability of the proposed approach to dealing large-scale classification problems under conditions of uncertainty.

This Paper is considered the pioneer paper that introducing the concept of the social machine learning under uncertainty conditions. The results of the comparison between TCML and DSCML prove the superiority of DSCML in terms of accuracy level and CPU time proved by T-TEST. The comparative results on six benchmarking problems are done indicate to highly improvement in the speed and accuracy of the solution.

References

- [1] Baum E. B. and David Haussler. What size nets give valid generalization. Princeton University, Princeton, NJ 08540 .*Neural Computation* 1, 151-160 @ 1989 Massachusetts Institute of Technology, 1989.
- [2] Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* 38 (2): 325–339. Doi:10.1214/aoms/1177698950, 1967.
- [3] Girosi F. Approximation error bounds that use VC-bounds. Appearing in *Proceedings of the International Conference on Artificial Neural Networks*, Paris. 295-302, 1995.
- [4] Joseph C. Giarratano and Gary D. Riley. Expert Systems: principles and programming. Ed. Thomson Course Tech., ISBN 0-534-38447-1, 2005.
- [5] Key, J. T., et al. Bayesian Model Choice: What and Why. (With discussion), in *Bayesian Statistics MA*; New York: Oxford University Press, (Vol. 6), pp. 343–370, 1999.
- [6] Kon, M. A., et al. Extending Girosi's approximation estimates for functions in Sobolev spaces via Statistical Learning Theory. *Journal of Analysis and Applications*, 3(2). 67-90, 2005.
- [7] Kon M. A. And Raphael L. A. Approximating functions in reproducing kernel Hilbert spaces via Statistical Learning Theory in Wavelets and Splines. 271 Nashboro Press, Nashville, TN. 271-286, 2006.
- [8] Marcin Owczarczuk. Support vector machines with two support vectors. Warszawa, Poland. Department of Applied Econometrics website Resource document at: <http://www.sgh.waw.pl/instytut/zes/wp/>. Accessed 27 June 2014, 2009.
- [9] Miller. A. J., Subset Selection in Regression. (2nd ed.), London. ISBN 1-58488-171-2, 2002.
- [10] Mendelson, S. A few notes on Statistical Learning Theory, in *Advanced Lectures on Machine Learning*. LNCS 2600, *Springer York: Chapman & Hall*. ISBN: 978-3-540-00529-2 (Print) 978-3-540-36434-4 (Online), 2003.
- [11] Olcay Taner Yilidiz et al. On the VC-Dimension of the Univariate Trees. Appearing in *Proceedings of the International Conference on Pattern Recognition Applications and Methods*. 193-198, 2012.
- [12] Onshuus, A. Usvyatsov. A. On dp-minimality, strong dependence and weight, the *journal of Symbolic Logic*. 76 (3), 737-740, 2011.
- [13] Rombaut, M. and Zhu, Y. Study on Dempster-Shafer theory for image segmentation applications. *Image and Vision Computing*, 20 (1): 15–23, 2002.
- [14] Sakurai, A. On the VC-dimension of neural networks with a large number of hidden layers appearing in *Proceedings of the International Conference Of NOLTA, IEICE*, 239-242, 1993.
- [15] Shafer, Glenn. A Mathematical Theory of Evidence. Princeton University Press, ISBN 0-608-02508-9, 1976.
- [16] Shelah. S. Strongly dependent theories. *Israel J. Math.*, to appear, available <http://arxiv.org/abs/math/0504197>, 2008.

- [17] Tazid Ali, et al. A New Combination Rule for Conflict Problem of Dempster-Shafer Evidence Theory. *International Journal of Energy, Information and Communications*.3 (1), 35-40, 2012.
- [18] Teytaud O. and LallichS. Contribution of statistical learning to validation of association rules. Extended version of "Bornesunformes en extraction de règlesd'association", in ***Proceedings of the International Conference ofCAP***,2001.
- [19] Vapnik V. and Chervonenkis A. A note on one class of perceptrons. *Automation and Remote Control*, 25,1964.
- [20] Vapnik V. andChervonenkis. Ordered risk minimization .*Remote Control*, 34:1226-1235,1974.
- [21] Vapnik V. ,Statistical Learning Theory. *New York: Wiley*,1998.
- [22]Vapnik.V.An overview of statistical learning theory. *IEEE Transaction on Neural Networks*, 10:988-999, 1999.
- [23] Zaher et al.A Social learning Approach for Minimizing True Risk ofCollective Machine Learning .*International Journal of Advanced Research in Computer Science and Software Engineering* 3(11), pp. 1172-1179,2013.