AMTEA-based Multi-task Optimisation for Multi-objective Feature Selection in Classification

Abstract. Feature selection is important nowadays due to many real-world datasets usually having a large number of features. Evolutionary multi-objective optimisation algorithms have been successfully used for feature selection which usually has two conflicting objectives, i.e., maximising the classification accuracy and minimising the number of selected features. However, most of the existing evolutionary multi-objective feature selection algorithms tend to address feature selection tasks independently, even when these feature selection tasks are related. Multi-task optimisation, which aims to improve the performance of multiple tasks by sharing common knowledge among them, has been used in many areas. However, there is not much work on utilising multi-task optimisation for feature selection. In this work, we develop a new multi-task multi-objective feature selection algorithm. This algorithm aims to address multiple related feature selection tasks simultaneously and facilitate knowledge capturing and transferring among the related tasks. Furthermore, a method is developed for transferring knowledge between related feature selection tasks having different features. This method can avoid transferring information between the unique features of tasks by transforming the probability models of them. We compare the proposed algorithm with the single-task multi-objective feature selection algorithm on seven sets of related feature selection tasks. Experimental results show that the proposed algorithm achieves better classification performance than the single-task algorithm with the help of knowledge transferring among related feature selection tasks. Further analysis shows that the features selected by our proposed algorithm can be more relevant to the classification tasks.

Keywords: feature selection · multi-task optimisation · multi-objective optimisation · evolutionary computation.

1 Introduction

Many real-world datasets contain a large number of features, which can be a challenge for classification. Feature selection which removes irrelevant and redundant features, can improve classification performance and speed up the learning process. Feature selection usually has two objectives, i.e. minimising the number of selected features and minimising the classification error rate. Most existing multi-objective feature selection algorithms tend to address feature selection tasks independently. However, many real-world feature selection tasks are related

and have common knowledge. For example, the IMDB movie review dataset and the MR movie review dataset have many common features about movie reviews, which indicates that feature selection on these two datasets are related and have common knowledge. Therefore, sharing common knowledge across the related feature selection tasks can potentially improve their performance.

Multi-task optimisation, which aims to address multiple related learning tasks simultaneously and transfer common knowledge across them has been used in many areas, such as the traveling salesman problem [14] and job-shop scheduling problems [21,22] and symbolic regression problems [20] in recent years. However, the research on multi-task optimisation for feature selection has not been much. In [4,5], multi-task optimisation was used for addressing high-dimensional feature selection tasks. In their work, an assistant feature selection task with a lower dimensionality is generated based on the main feature selection task. During the evolutionary process, knowledge is transferred from the assistant task to the main task. By doing this, the performance of the main task can be improved. One of the limitations of these two algorithms is that they only focus on the performance of the main task. However, multi-task optimisation aims to improve the performance of all the tasks. Furthermore, using existing multi-task optimisation algorithms for feature selection is still challenging. Many existing methods are based on transferring a number of solutions across multiple related tasks [8, 11]. The selection of transferred solutions is usually based on learning the relationships between tasks, which is usually achieved by analyzing the relatedness between the solutions of the multiple tasks, which can be challenging for feature selection.

An adaptive model-based transfer-enabled evolutionary algorithm (AMTEA) [6], which is an evolutionary sequential transfer learning, was proposed to address a target task with the help of knowledge learned from the experience of addressing the other tasks from the source domains. In AMTEA, analyzing the relationship between tasks is achieved by analyzing the relationships between the probabilistic distributions of solutions of them, which can be effective and efficient for feature selection. Multi-task optimisation can be treated as conducting multiple evolutionary sequential transfer learning simultaneously. Therefore, in this work, we aim to develop a new multi-task otpimisation algorithm based on AMTEA for feature selection. Different from AMTEA, our new multi-task optimisation algorithm focuses on addressing multiple related feature selection tasks simultaneously rather than addressing only one main task at a time. In AMTEA, an offline knowledge pool containing probabilistic models from the source domain is built for a target task. However, for our multi-task optimisation algorithm, we propose a method for building an online knowledge pool containing probabilistic models for multiple tasks. Although AMTEA can be useful for feature selection, it is challenging for AMTEA to deal with case feature selection tasks that have different features. In AMTEA, the search spaces of the source tasks and target tasks are transformed into a unified search space with the range of [0, 1]. In this way, information may be transferred between unique features of different tasks, which may not be effective for feature selection

since the meanings of different features are different. To address this challenge, we propose a strategy for transforming the probabilistic models when the related feature selection tasks have different features.

The main objectives of this paper can be summarised as follows.

- Develop a new multi-task optimisation algorithm for feature selection, which addresses multiple feature selection tasks simultaneously rather than just focusing on the main task at a time.
- Develop a method for building an online knowledge pool that provides knowledge for multiple tasks simultaneously.
- Develop a method for transforming the probabilistic models for the case multiple tasks have different features.
- Examine the effectiveness of the developed algorithm by comparing it with a single-task feature selection algorithm and testing on a set of benchmark datasets having different features.

2 Related Work

2.1 Evolutionary Multi-objective Feature Selection

In recent years, evolutionary multi-objective optimisation has achieved great success in feature selection. Recently, a new multi-objective wrapper method was developed for feature selection in image classification [10]. In their work, new representations and breeding operators are developed to enhance the performance of multi-objective feature selection methods. Wang et al. [17] develop an enhanced multi-objective feature selection algorithm based on the sampling strategy to reduce the computational cost and improve the performance. In their work, K-means clustering based on differential selection and a ladder-like sample utilization strategy is proposed to reduce the size of training samples, improving feature selection's efficiency. In summary, many of these algorithms have shown their effectiveness in solving feature selection tasks. However, they tend to address each feature selection task separately, while many real-world feature selection tasks connect to each other and share common knowledge for problem-solving. It is worth exploring methods to utilize this knowledge to enhance the related feature selection tasks.

2.2 Multi-task Optimisation for Feature Selection

Multi-task optimisation, which aims to address multiple related optimistation tasks simultaneously and facilitate knowledge transfer across them, has been successfully used in many areas in recent years. However, research on multi-task optimisation for feature selection has still been limited. In [4], a multi-task optimisation-based algorithm is proposed to address high-dimensional feature selection problems. In their work, a knee point selection scheme is developed to distinguish promising features from all features. An assisted task is generated by selecting a subset of features from the target task. The target task and

the assisted task are addressed simultaneously, and a new crossover operator is proposed to transfer information. It has been shown that the performance on the target task can be significantly improved with the help of information from the assisted task. One of the limitations of their work is that their method only focuses on the performance of the main feature selection task rather than the performance of all the tasks. Although these recently proposed algorithms have shown the potential of multi-task optimisation for feature selection, it is still challenging to use existing algorithms for feature selection. The main reason is that many existing multi-task optimisation algorithms are based on transferring a number of solutions across multiple related tasks [8, 11]. In these existing algorithms, transferring useful solutions across tasks helps achieve positive transfer. The selection of transferred solutions is usually based on learning the relationships between tasks, which is achieved by learning the relatedness between the current populations of these tasks. However, feature selection tasks usually have many features, making it hard to analyze the relationship between tasks in feature selection. Furthermore, the complex interactions between features can also make it challenging to the relationships between features selection tasks.

2.3 Adaptive Model-based Transfer-enabled Evolutionary Algorithm–AMTEA

The adaptive model-based transfer-enabled evolutionary algorithm (AMTEA), an evolutionary sequential transfer learning algorithm, was proposed in [6]. AMTEA aims to improve the performance on a target task with the help of knowledge from source tasks that have been addressed. Assume the total number of target and source tasks is K, including one target task and $K-1$ source tasks. In AMTEA, a probabilistic model is used to describe the solution distribution of one task. Thus, the knowledge for solving the task is stored in the probabilistic model. Compared with storing all the solutions, storing probabilistic models can save a huge amount of computational memory. $K-1$ probabilistic models for the $K-1$ addressed (source) tasks are built at the beginning of the algorithm. Then, at each generation, a probabilistic model is built to describe the solution distribution of the target task. Furthermore, a mixture probabilistic model is built based on the combination of the K models, including the target and $K-1$ source models. The mixture model is built based on the relatedness between the source and the target tasks. If the source task is highly related to the target task, its probabilistic model will greatly impact the mixture model. Compared with learning the relationships between the solutions of different tasks, learning the relationships between the probabilistic models of different tasks can be more effective and efficient. After the mixture model is built, new solutions will be generated based on it and transferred to the target task. By doing this, knowledge for problem-solving is transferred from the source tasks to the target task.

AMTEA has a good potential to be developed as a multi-task optimisation algorithm for feature selection. However, it is still challenging. Firstly, AMTEA is based on the assumption that all the source tasks have been addressed. A

knowledge pool containing probability models is built based on the solutions of the source tasks. However, in multi-task optimisation, all the tasks are addressed simultaneously, which makes it challenging to build a knowledge pool for them. To address this issue, a method for building an online knowledge pool for knowledge transferring among multiple tasks simultaneously is developed. Furthermore, it is challenging for AMTEA to transfer knowledge between feature selections having different features. Using the transferring strategy in AMTEA, information may be transferred between the unique features from different tasks, which does not work for feature selection since different features have different meanings. To address this issue, a method for transferring knowledge across related feature selection tasks having different features is proposed. The proposed algorithm is expected to achieve better performance than state-of-the-art feature selection algorithms without significantly increasing computational time.

3 Proposed Method

In this work, we develop a new multi-task multi-objective feature selection algorithm based on AMTEA named FSMTO, which includes two main components: first, a method to build an online knowledge pool storing the probabilistic models which describe the distributions of solutions in multiple tasks during the evolutionary process; second, a method for transforming probabilistic models for the case tasks have different features. This algorithm aims to address multiple related feature selection tasks having common features simultaneously. The online knowledge pool is used for storing the captured knowledge of addressing the multiple related tasks during the evolutionary process. In this paper, knowledge is transferred across the common features of the multiple related feature selection tasks. Therefore, the probabilistic models are transformed when the multiple tasks have different features. This framework can work with many evolutionary multi-objective feature selection algorithms such as NSGA-II [7], IBEA [24], and MOEA/D [23]. In the following sections, the framework of the proposed algorithm is described first. Then, the details of the two components will be introduced.

3.1 The Framework of FSMTO

Fig. 1 shows the workflow of the proposed algorithm. As shown in Fig. 1, K feature selection tasks represented by $T_k, k = 1, 2, ..., K$ are addressed simultaneously. An online knowledge pool containing K probabilistic models represented by $M_k^t, k = 1, 2, ..., K$ is built for the multiple tasks at the t^{th} generation. At the $t-1^{th}(t > 0)$ generation, K probabilistic models are built based on the populations of the K feature selection tasks. Then, knowledge is transferred from the knowledge pool to each of the K feature selection tasks at the t^{th} generation by using the knowledge transferring method in AMTEA.

Algorithm 1 shows the pseudocode of the proposed multi-task multi-objective feature selection algorithm FSMTO. As can be seen, K populations are randomly

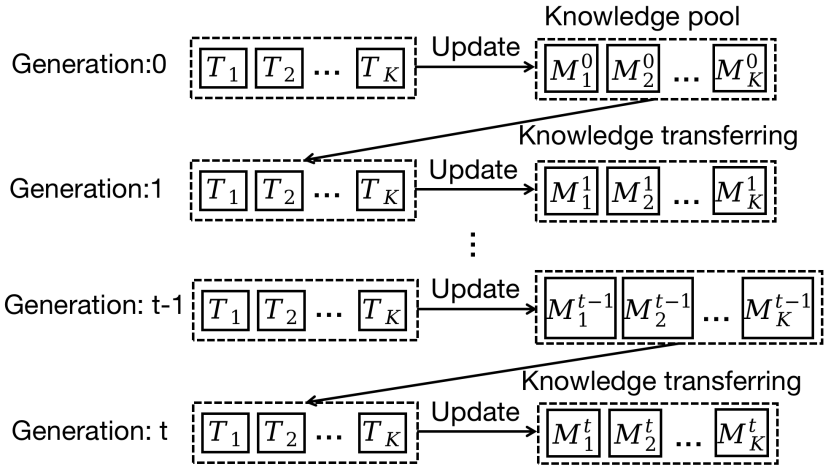

Fig. 1. The workflow of the proposed algorithm FSMTO.

initialized for the K tasks at the beginning of FSMTO. Each population has a size of N. During the evolutionary process, new solutions for each task are generated either based on knowledge transfer across tasks or genetic operators such as cross-over and mutation, which is controlled by the transfer interval Δ between the two closest knowledge transfers. When Δ is large, most of the new solutions for each task are generated by genetic operators. Otherwise, most of the new solutions are generated based on knowledge transfer. After the new solutions are generated, a set of solutions with better objective values will be selected for the next generation under our proposed framework. Finally, the proposed algorithm will select K sets of solutions (feature subsets) for the K feature selection tasks.

As mentioned previously, the key idea of AMTEA is to build a mixture model based on the source probabilistic models for the target task. Since FSMTO is a multi-task optimisation algorithm, knowledge transferring based on AMTEA is conducted K times simultaneously for K feature selection tasks. Here, we introduce how to transfer knowledge from the knowledge pool to one of the K feature selection tasks. The rest $K-1$ tasks follow the same way. Following AMTEA, a mixture model is built to approximate the distribution of optimal solutions for the target task. Let FS represent a set of optimal solutions of the feature selection task, $P^*(\mathbf{x})$ represents the probabilistic distribution of FS. The $K-1$ source probabilistic models are used to approximate the true latent probabilistic model $P^*(\mathbf{x})$ of the target task. At the t^{th} generation, the approximated true latent probabilistic model for the target task is represented by $Q_t(\mathbf{x})$. Let ϕ_K represent the probabilistic model built based on the population of the target task at the t^th generation. They satisfy the following relationship.

$$Q_t(\mathbf{x}) \approx \Sigma_{l=1}^{K} \alpha_l^t \phi_l(\mathbf{x}) \tag{1}$$

where $\phi_l(\mathbf{x}), l = 1, 2, ..., K-1$ are the $K-1$ source probabilistic models. α_l^t represents the coefficient of the l^{th} model. At the t^{th} generation, the current population of the target task can be represented by U^t. To approximate $Q_t(\mathbf{x})$, the α_l^ts that maximise the probability of observing the data U_t should be found. Since $Q_t(\mathbf{x})$ is used to approximate $P^*(\mathbf{x})$, the following relationship will be satisfied.

$$P^*(\mathbf{x}) \approx lim_{t \to +\infty} Q_t(\mathbf{x}) \tag{2}$$

where, $q(\mathbf{x}|t)$ represents the approximation of $P_t(\mathbf{x})$ at the t^{th} generation. $\phi_l(\mathbf{x}), l = 1, 2, ..., K-1$ represent the $K-1$ probabilistic models of the source tasks, while $\phi_K(\mathbf{x})$ is a model describing the target dataset U_t. Obtaining α_l^t of the mixture model is equal to maximising Eq. (3).

$$\log L = \Sigma_{i=1}^{N} \log \Sigma_{l=1}^{K} \alpha_l^t \phi_l(\mathbf{x}) \tag{3}$$

where, N is the number of instances of U_t. The log function helps to transform the products in the objective into sums, which is easier to deal with. $\mathbf{x} \in U_t$. Transfer coefficients α_l^t are related to the similarity between the solution distributions of the source task and the target task. The value of α_l^t is in the range of [0,1]. If α_l^t is close to 1, the l^{th} source task is similar to the target task.

After the k^{th} mixture model is built, new solutions are sampled based on the k^{th} mixture model and transferred to the k^{th} feature selection task, which achieves knowledge transferring.

3.2 Build An Online Knowledge Pool

In FSMTO, multiple feature selection tasks are addressed parallelly at the same time. At each generation of the evolutionary process, knowledge of solving each task can be captured from the population and stored in a probabilistic model and then shared across tasks. An online knowledge pool that contains these probabilistic models is built. The probability model used in this work is based on the assumption that features are independent. K probabilistic models are built for the K feature selection tasks at each generation.

The probabilistic model, which describes the current population of a feature selection task, is shown as follows.

$$P(o_1 = k_{o_1}, o_2 = k_{o_2}, ..., o_D = k_{o_D}) = \prod_{i=1}^{D} p_i^{k_{o_i}} (1 - p_i)^{1 - k_{o_i}} \tag{4}$$

where, o_i represents whether the ith feature is selected. $k_{o_i} \in \{0, 1\}$. p_i is the probability of selecting the ith feature, which can be calculated based on the solutions of the feature selection task. For example, if the population of a feature selection task consists of $c_1 = [1, 1, 0, 0], c_2 = [1, 0, 0, 0], c_3 = [0, 1, 0, 1]$. Then, the probabilistic model is:

$$\begin{aligned} &P(o_1 = k_{o_1}, o_2 = k_{o_2}, o_3 = k_{o_3}, o_4 = k_{o_4}) = \\ &((\frac{2}{3})^{k_{o_1}} (\frac{1}{3})^{1-k_{o_1}})(\frac{2}{3}^{k_{o_2}} \frac{1}{3}^{1-k_{o_2}})(0^{k_{o_3}} 1^{1-k_{o_3}})((\frac{1}{3})^{k_{o_4}} (\frac{2}{3})^{1-k_{o_4}}) \end{aligned} \tag{5}$$

Algorithm 1 Pseudocode of FSMTO

Input:
K related feature selection tasks and a transfer interval Δ;
Output:
Solution(s) to all the K feature selection tasks;
Set $t = 1$;
Generate K initial populations $P_1(t), ..., P_K(t)$ for the K tasks, the size of each population is N;
Evaluate the individuals based on their objective functions
while t is not larger than the maximum generation **do**
  Update the probabilistic models in the knowledge pool based on the K current populations.
  if mod (t+1, Δ) == 0 **then**
    for k=1:K **do**
      Build a mixture model $q_k(\mathbf{x}|t)$ for the k^{th} task according to Eq (3);
      Randomly sample the offspring population $P_k^c(t)$ from the mixture model $q_k(\mathbf{x}|t)$;
  else
    for k=1:K **do**
      Randomly select individuals from $P_k(t)$ to form a parent population $P_k^s(t)$;
      Use genetic operators including crossover and mutation to generate an offspring population $P_k^c(t)$ based on the parent population $P_k^s(t)$;
  for k=1:K **do**
    Evaluate individuals in $P_k^c(t)$ based on their objective functions;
    Select next generation $P_k(t+1)$ from $P_k(t) \cup P_k^c(t)$ based on multi-objective optimisation algorithms;
  Set $t = t + 1$;

At the current generation, K newly generated probabilistic models replace the K old models and update the knowledge pool.

3.3 Transform the Probabilistic Models

When the search spaces of related tasks are different, the probabilistic model of one task can not be transferred to another task without any transformation. Since different features have different meanings, knowledge needs to be transferred across common features of the related feature selection tasks. In this work, each task will be treated as both the target task and the source task; thus knowledge transferring is conducted K times among these tasks.

We introduce how to transform the probabilistic model of a source task. The rest tasks follow the same way. Let $V_{target} = \{o_1^t, ..., o_a^t, u_1^t, ..., u_b^t\}$ represent the feature vector of a target feature selection task, while $V_{source} = \{o_1^s, ..., o_a^s, z_1^s, ..., z_c^s\}$ represents the feature vector of a source feature selection task where $o_i^t = o_i^s, i = 1, ..., a$. Here $o_1^t, ..., o_a^t$ and $o_1^s, ..., o_a^s$ represent the common features of the target and the source tasks. $u_1^t, ..., u_b^t$ and $z_1^s, ..., z_c^s$ represent the unique features from the target task and the source task, respectively. In this case, the target probabilistic model can be split into two parts as shown in

Eq. (6), where the first part is $P(o_1^t = k_{o_1^t}, ..., o_a^t = k_{o_a^t}) = \prod_{i=1}^{a} {p_{o_i^t}}^{k_{o_i^t}} (1 - p_{o_i^t})^{1-k_{o_i^t}}$ and the second part is $P(u_1^t = k_{u_1^t}, ..., u_c^t = k_{u_c^t}) = \prod_{i=1}^{c} {p_{u_i^t}}^{k_{u_i^t}} (1 - p_{u_i^t})^{1-k_{u_i^t}}$.

$$\begin{aligned} P(o_1^t = k_{o_1^t}, ..., o_a^t = k_{o_a^t}, u_1^t = k_{u_1^t}, ..., u_c^t = k_{u_c^t}) = \\ \prod_{i=1}^{a} {p_{o_i^t}}^{k_{o_i^t}} (1 - p_{o_i^t})^{1-k_{o_i^t}} \cdot \prod_{i=1}^{c} {p_{u_i^t}}^{k_{u_i^t}} (1 - p_{u_i^t})^{1-k_{u_i^t}} \end{aligned} \tag{6}$$

Similarly, the source probabilistic model is represented by

$$\begin{aligned} P(o_1^s = k_{o_1^s}, ..., o_a^s = k_{o_a^s}, z_1^s = k_{z_1^s}, ..., z_c^s = k_{z_c^s}) = \\ \prod_{i=1}^{a} {p_{o_i^s}}^{k_{o_i^s}} (1 - p_{o_i^s})^{1-k_{o_i^s}} \cdot \prod_{i=1}^{c} {p_{z_i^s}}^{k_{z_i^s}} (1 - p_{z_i^s})^{1-k_{z_i^s}} \end{aligned} \tag{7}$$

As can be seen, the differentes between the target probabilistic model and the source probabilistic model are

$$P(u_1^t = k_{u_1^t}, ..., u_c^t = k_{u_c^t}) = \prod_{i=1}^{c} {p_{u_i^t}}^{k_{u_i^t}} (1 - p_{u_i^t})^{1-k_{u_i^t}} \tag{8}$$

and

$$P(z_1^s = k_{z_1^s}, ..., z_c^s = k_{z_c^s}) = \prod_{i=1}^{c} {p_{z_i^s}}^{k_{z_i^s}} (1 - p_{z_i^s})^{1-k_{z_i^s}} \tag{9}$$

Since features $z_1^s, ..., z_c^s$ are not in the target probabilistic model, Eq. (9) of the source probabilistic model should be replaced when the source probabilistic model is transferred to the target task. Due to the fact that the source task does not contain the information about the unique features in the target task $u_1^t, ..., u_b^t$, The probabilities of selecting features $u_1^t, ..., u_b^t$ in the source tasks are all 0. Eq. (9) of the source task will be replaced with

$$P(u_1^t = k_{u_1^t}, ..., u_b^t = k_{u_b^t}) = \prod_{i=1}^{b} 0^{k_{u_i^t}} 1^{1-k_{u_i^t}} \tag{10}$$

The transformed source probabilistic model which can be used for the target task is represented as follows.

$$\begin{aligned} P(o_1^s = k_{o_1^s}, ..., o_a^s = k_{o_a^s}, u_1^t = k_{u_1^t}, ..., u_b^t = k_{u_b^t}) = \\ \prod_{i=1}^{a} {p_{o_i^s}}^{k_{o_i^s}} (1 - p_{o_i^s})^{1-k_{o_i^s}} \cdot \prod_{i=1}^{b} 0^{k_{u_i^t}} 1^{1-k_{u_i^t}} \end{aligned} \tag{11}$$

4 Experiment Design

In this paper, improving the strength pareto evolutionary algorithm (SPEA2), which is the commonly used multi-objective framework [24], is used as a baseline. We compare SPEA2 with SPEA2 cooperating with the proposed algorithm, named SPEA2-FSMTO. Firstly, we compare the hypervolume values obtained by the two algorithms on the training sets of the benchmark datasets. Secondly, we compare the hypervolume values obtained by the two algorithms on the test sets of the benchmark datasets. Furthermore, we compare the computational time cost of the two algorithms on the benchmark datasets. Finally, further analysis of the selected features of the two algorithms is conducted.

Table 1. Datasets

Dataset	Instance	Features	Class	Dataset	Instance	Features	Class
Wine-1	1599	12	2	Dermatology-1	358	34	2
Wine-2	4898			Dermatology-2			
Mushroom-1	3516	22	2	Dermatology-3			
Mushroom-2	4608			Dermatology-4			
Magic-1	2212	10	2	Dermatology-5			
Magic-2	16808			Dermatology-6			
Letter-1	1151	16	2	Waveform-1	1331	40	3
Letter-2	1543			Waveform-2	3669		
Letter-3	1509			News-1	3749	200	4
Letter-4	1575			News-2	3743		
Letter-5	1536			News-3	3702		
Letter-6	1573			News-4	3669		

4.1 Benchmark Datasets

Multiple related datasets, which have common features but different distributions, and similar tasks, have been used in this paper. The feature selection tasks on these datasets are related and have common knowledge, which can be used for evaluating the performance of our proposed algorithm. Their properties, such as the number of features, instances, and classes of these datasets, are summarised in Table 1.

- **Wine**: The Red Wine dataset and the White Wine dataset, which have the same features but different data distributions, are the two representative datasets in the field of wine quality classification [2].
- **Mushroom**: Mushroom is used to classify whether a mushroom is edible or poisonous. Two related datasets generated from the original Mushroom dataset are widely used in transfer learning [15].
- **Magic**: The Magic dataset contains image data of hadronic showers. This dataset is used for the classification of hadronic showers caused by primary gammas or upper atmospheres. The original dataset can be divided into two related sub-datasets with the same feature space, and different data distributions [16].
- **Letter**: Letter Recognition dataset contains image data of letters from A to Z [9]. This dataset also can be divided into several binary classification datasets, which have the same search space but different data distributions. Based on this, six sub-datasets '**I** vs **T**', '**E** vs **F**', '**C** vs **G**', '**M** vs **N**', '**Q** vs **O**' , and '**X** vs **Y**' are generated, respectively. '**I** vs **T**' denotes as a sub-dataset with the task of classification between letter '**I**' and letter '**T**', which contains the instances with a label of '**I**' or '**T**'. This also applies for the other five sub-datasets.
- **Dermatology**: Dermatology contains instances of six dermatological diseases. This dataset has 33 clinical and histopathological attributes. Following [1], Dermatology can be divided into six sub-datasets which are used for binary classification tasks. These sub-datasets have the same search space but different data distributions, which can be used for transfer learning.

- **Waveform**: The Waveform dataset, which is usually used for transfer learning, contains the data of three classes of waves. This original dataset can be transformed into two related datasets where Waveform-1 contains the instances which have the first feature larger than 0.15 and the second feature larger than 0, and Waveform-2 contains the remaining instances.
- **20 Newsgroups**: The 20 Newsgroups dataset contains nearly 20,000 newsgroup documents, evenly across 20 different news categories [3]. In this research, we create four related sub-datasets based on the original dataset, which are named News-1, News-2, News-3, and News-4. The properties of these four generated sub-datasets, including the document categories, are summarised in Table 2. As can be seen, the search spaces of the feature selection tasks are similar but different, which is sensible to evaluate the effectiveness of our proposed method for transforming the probabilistic models across tasks having different features.

Table 2. Document Categories in News Datasets

Name	Documents Categories	Name	Documents Categories
News-1	alt.atheism,comp.graphics, rec.autos,sci.space	News-3	alt.atheism,soc.religion.christian, talk.politics.guns,rec.motorcycles
News-2	alt.atheism,comp.graphics, misc.forsale,rec.motorcycles	News-4	alt.atheism,rec.sport.hockey, talk.politics.guns,sci.crypt

4.2 Parameter Settings

In the experiments, each dataset is randomly split with 70 % forms the training set, while the rest 30% will be put into the test set. The classification error rate, the second objective value of multi-objective feature selection, of candidate feature subsets is calculated based on the Decision Tree with 10-fold cross-validation on the training set [13].

Following [12], in this work, the population size is set to the number of features in each task. For the purpose of having enough computational resources for the feature selection algorithms to find a set of good solutions, the maximal number of generations is set to five times the population size but will not be larger than 200. Each algorithm will run 30 independent times on each dataset. The one-point crossover operator and mutation are used in this research. The probability of mutation is set to $1/D$, where D is the number of features [12]. The transfer interval of AMTEA is set to 2. Table 3 shows the summary of the parameter settings of this paper.

4.3 Result Analyses and Discussions

In this section, the comparisons between the performance of SPEA2-FSMTO and SPEA2 are used to verify the effectiveness of the proposed multi-task optimisation algorithm. Hypervolume (HV) [18], which is one of the most popular

Table 3. The parameter setting

Parameters	Settings
Population size	N
Maximum generation	Min(200, 5*N)
Dimension of solutions	D
Probability of mutation	1/D
Transfer interval	2
Reference point for hypervolume	(1.1,1.1)

Table 4. Hypervolume on the training sets

	SPEA2	SPEA2-FSMTO			SPEA2	SPEA2-FSMTO	
	Mean(Std)	Mean(Std)	W-test		Mean(Std)	Mean(Std)	W-test
Wine-1	0.956(0.033)	**0.976(0.004)**	+	Dermatology-1	1.171(0.003)	**1.173(0.001)**	+
Wine-2	0.909(0.024)	**0.919(0.003)**	+	Dermatology-2	1.15(0.011)	1.153(0.003)	≈
Mushroom-1	1.153(0.009)	**1.157(0.0)**	+	Dermatology-3	1.178(0.0)	**1.178(0.0)**	+
Mushroom-2	1.159(0.002)	**1.16(0.0)**	+	Dermatology-4	1.159(0.003)	**1.159(0.002)**	+
Magic-1	0.972(0.037)	**0.987(0.023)**	+	Dermatology-5	1.176(0.003)	1.176(0.003)	≈
Magic-2	0.867(0.035)	**0.873(0.032)**	+	Dermatology-6	1.175(0.002)	**1.177(0.001)**	+
Letter-1	1.091(0.036)	**1.122(0.005)**	+	Waveform-1	0.921(0.01)	**0.927(0.009)**	+
Letter-2	1.116(0.024)	**1.13(0.002)**	+	Waveform-2	0.903(0.02)	**0.914(0.01)**	+
Letter-3	1.053(0.028)	**1.073(0.005)**	+	News-1	0.789(0.016)	**0.909(0.005)**	+
Letter-4	1.078(0.039)	**1.104(0.006)**	+	News-2	0.814(0.016)	**0.94(0.004)**	+
Letter-5	1.037(0.06)	**1.082(0.004)**	+	News-3	0.779(0.015)	**0.882(0.006)**	+
Letter-6	1.113(0.023)	**1.127(0.001)**	+	News-4	0.82(0.02)	**0.94(0.005)**	+

performance indicators in evolutionary multi-objective optimisation is used for comparisons. A larger HV value indicates better multi-objective optimisation performance. Since the two objectives of feature selection are both in the range of [0,1], the reference point of HV is set as (1.1, 1.1) in this paper [19]. The Wilcoxon test with a significance level of 0.05 is used for the performance comparison in this paper. The sign of '+' means that SPEA2-FSMTO is significantly better than SPEA2, while the sign of '−' means that SPEA2-FSMTO is significantly worse than SPEA2. The sign of '≈' means that SPEA2-FSMTO ties SPEA2. Finally, the features selected by the two compared algorithms on one typical dataset, Magic, are analyzed.

Comparisons on Hypervolume Values on the Training Sets and the Test Sets: Tables 4 and 5 show the averages and standard deviations of the HV values calculated based on the final fronts obtained by the two algorithms, SPEA2-FSMTO and SPEA2, over the 30 independent runs on the training sets and test sets. Based on the statistical test results, SPEA2-FSMTO has a significantly better training performance than SPEA2 on 22 of the 24 benchmark datasets. Furthermore, the proposed algorithm has a better testing performance than SPEA2 on 18 of the 24 tasks. Overall, with the help of knowledge transferring among related feature selection tasks, the proposed algorithm can obtain a better performance than SPEA2 on the benchmark datasets.

Table 5. Hypervolume on the test sets

	SPEA2	SPEA2-FSMTO			SPEA2	SPEA2-FSMTO	
	Mean(Std)	Mean(Std)	W-test		Mean(Std)	Mean(Std)	W-test
Wine-1	0.951(0.036)	**0.972(0.008)**	+	Dermatology-1	1.176(0.003)	1.176(0.003)	≈
Wine-2	0.909(0.025)	0.919(0.004)	≈	Dermatology-2	1.114(0.017)	1.119(0.007)	≈
Mushroom-1	1.153(0.009)	**1.157(0.0)**	+	Dermatology-3	**1.16(0.006)**	1.158(0.0)	−
Mushroom-2	1.159(0.003)	**1.16(0.0)**	+	Dermatology-4	**1.152(0.007)**	1.145(0.009)	−
Magic-1	0.956(0.036)	**0.973(0.023)**	+	Dermatology-5	1.178(0.0)	1.178(0.0)	≈
Magic-2	0.868(0.035)	**0.874(0.032)**	+	Dermatology-6	1.173(0.006)	**1.175(0.004)**	+
Letter-1	1.091(0.036)	**1.122(0.005)**	+	Waveform-1	0.896(0.013)	0.902(0.019)	≈
Letter-2	1.116(0.024)	**1.13(0.002)**	+	Waveform-2	0.89(0.022)	**0.901(0.014)**	+
Letter-3	1.053(0.028)	**1.073(0.005)**	+	News-1	0.74(0.015)	**0.869(0.007)**	+
Letter-4	1.078(0.039)	**1.104(0.006)**	+	News-2	0.785(0.015)	**0.925(0.006)**	+
Letter-5	1.037(0.06)	**1.082(0.004)**	+	News-3	0.695(0.017)	**0.796(0.01)**	+
Letter-6	1.111(0.023)	**1.125(0.003)**	+	News-4	0.758(0.019)	**0.875(0.007)**	+

Table 6. Computational time (in Seconds) of the algorithms on the training sets

	SPEA2	SPEA2-FSMTO	
	Mean(Std)	Mean(Std)	W-test
Wine	51.296(3.068)	**47.048(2.474)**	+
Mushroom	126.609(2.897)	**113.095(3.624)**	+
Magic	139.418(11.916)	**117.064(5.654)**	+
Letter	207.59(24.883)	**180.798(2.852)**	+
Dermatology	794.868(22.465)	**690.16(5.253)**	+
Waveform	805.292(41.735)	**671.073(25.318)**	+
News	5679.773(663.745)	5461.877(1344.209)	≈

Table 7. Features selected by SPEA2 and SPEA2-FSMTO on Magic-1 dataset.

SPEA2 on Magic-1	
Classification accuracy	Selected features
0.899844994	fLength, fSize, fConc, fAsym, fM3Long, fAlpha
0.896606619	fLength, fM3Long, fAlpha
0.860410557	fM3Long, fAlpha
SPEA2-FSMTO on Magic-1	
Classification accuracy	Selected features
0.914046921	fLength, fWidth, fSize, fM3Long, fAlpha
0.896606619	fLength, fM3Long, fAlpha
0.873338919	fLength, fAlpha
0.757246025	fLength, fWidth, fAlpha

Comparisons on Computational Time: To investigate the efficiency of the proposed feature selection framework, we also compare the computational cost of the two feature selection algorithms on the training sets. Table 6 shows the average and standard deviation of the computational time (in seconds) of the two algorithms over 30 runs. The results show that SPEA2-FSMTO spends less computational time than SPEA2 for feature selection. There are several possible reasons: first, building the mixture model in AMTEA does not consume many computational resources [6]; second, nearly half of the new solutions are generated by probabilistic models in SPEA2-FSMTO, which may be faster

Table 8. Features selected by SPEA2 and SPEA2-FSMTO on Magic-2 dataset.

SPEA2 on Magic-2	
Classification accuracy	Selected features
0.796940798	fWidth, fSize, fConc, fAlpha, fDist
0.790144579	fLength, fWidth, fSize, fAlpha
0.757246025	fLength, fWidth, fAlpha
SPEA2-FSMTO on Magic-2	
Classification accuracy	Selected features
0.797450569	fLength, fWidth, fSize, fM3Long, fAlpha
0.790144579	fLength, fWidth, fSize, fAlpha
0.765918555	fWidth, fSize, fAlpha
0.757246025	fLength, fWidth, fAlpha

than generating all the new solutions by genetic operators in SPEA2; Finally, SPEA2-FSMTO selects a smaller number of features than SPEA2 (see the next subsection), which helps to save computational cost in fitness evaluations. Therefore, FSMTO can improve the performance of SPEA2 with better efficiency for most datasets.

Further Analysis on the Selected Features: To further understand the effectiveness of the proposed method, in this subsection, we compare the features selected by the single-task multi-objective feature selection algorithm and the proposed multi-task multi-objective feature selection algorithm. The Magic datasets contain ten features, which is feasible for our analysis. Furthermore, the meaning of each feature of the Magic datasets is known, which is helpful in understanding why the selected features result in good classification performance. Therefore, we use the two Magic datasets as an example for our analysis. Table 7 and Table 8 show the features selected by the two algorithms on the run that has the median hypervolume values. As can be seen, SPEA2-FSMTO obtains better classification performance and selects a smaller number of features than SPEA2. SPEA2 selects features including {*fLength, fSize, fCon, fAsym, fM3Long, fAlpha*} and obtain a classification accuracy of 0.899 on Magic-1, while SPEA2-FSMTO selects features {*fLength, fWidth, fSize, fM3Long, fAlpha*} and obtains a classification accuracy of 0.914 on Magic-1. The main difference between them is that SPEA2-FSMTO selects the feature {*fWidth*}, but SPEA2 does not, which means the feature {*fWidth*} is useful to the classification task of the Magic-1 dataset. Furthermore, both SPEA2 and SPEA2-FSMTO select the feature {*fWidth*} on the Magic-2 dataset. Since SPEA2-FSMTO addressed the two related feature selection tasks, Magic-1 and Magic-2, together, the knowledge of selecting the feature {*fWidth*} can be transferred across them. However, the knowledge of selecting the feature {*fWidth*} can not be transferred from the feature selection task Magic-2 to Magic-1 in SPEA2, which results in lower classification accuracy on Magic-1 than SPEA2-FSMTO. Overall, SPEA2-FSMTO can select good features and achieve improved classification performance with the help of knowledge transfer across related feature selection tasks.

5 Conclusions

In this work, a new multi-task optimisation framework is developed for feature selection based on AMTEA. The proposed algorithm addresses multiple related feature selection tasks simultaneously rather than addressing one main task at a time. A method is developed to build an online knowledge pool that contains knowledge for multiple tasks. Furthermore, a method is developed to transform the probabilistic models for the case tasks with different features. To investigate the effectiveness of the proposed algorithm, related feature selection tasks having different features have been tested. Experimental results show that SPEA2 co-operating with the proposed algorithm selects a smaller number of features and achieves a better classification performance than that of SPEA2 based feature selection without knowledge transfer across tasks. Furthermore, the experimental results verified the effectiveness of the proposed for addressing multiple tasks having different features simultaneously.

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine learning **73**(3), 243–272 (2008)
2. Cao, B., Pan, S.J., Zhang, Y., Yeung, D.Y., Yang, Q.: Adaptive transfer learning. In: proceedings of the AAAI Conference on Artificial Intelligence. vol. 24 (2010)
3. Chandra, A.: Comparison of feature selection for imbalance text datasets. In: 2019 International Conference on Information Management and Technology (ICIMTech). vol. 1, pp. 68–72. IEEE (2019)
4. Chen, K., Xue, B., Zhang, M., Zhou, F.: An evolutionary multitasking-based feature selection method for high-dimensional classification. IEEE Transactions on Cybernetics **52**, 7172 – 7186 (2020)
5. Chen, K., Xue, B., Zhang, M., Zhou, F.: Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimisation. IEEE Transactions on Evolutionary Computation **26**, 446 – 460 (2021)
6. Da, B., Gupta, A., Ong, Y.S.: Curbing negative influences online for seamless transfer evolutionary optimization. IEEE Transactions on cybernetics **49**(12), 4365–4378 (2018)
7. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Transactions on evolutionary computation **6**(2), 182–197 (2002)
8. Feng, L., Huang, Y., Zhou, L., Zhong, J., Gupta, A., Tang, K., Tan, K.C.: Explicit evolutionary multitasking for combinatorial optimization: A case study on capacitated vehicle routing problem. IEEE transactions on cybernetics **51**(6), 3143–3156 (2020)
9. Gonçalves, A.R., Das, P., Chatterjee, S., Sivakumar, V., Von Zuben, F.J., Banerjee, A.: Multi-task sparse structure learning. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 451–460 (2014)
10. González, J., Ortega, J., Damas, M., Martín-Smith, P., Gan, J.Q.: A new multi-objective wrapper method for feature selection–accuracy and stability analysis for bci. Neurocomputing **333**, 407–418 (2019)

11. Lin, J., Liu, H.L., Xue, B., Zhang, M., Gu, F.: Multiobjective multitasking optimization based on incremental learning. IEEE Transactions on Evolutionary Computation **24**(5), 824–838 (2019)
12. Nguyen, B.H., Xue, B., Andreae, P., Ishibuchi, H., Zhang, M.: Multiple reference points-based decomposition for multiobjective feature selection in classification: Static and dynamic mechanisms. IEEE Transactions on Evolutionary Computation **24**(1), 170–184 (2019)
13. Nguyen, B.H., Xue, B., Zhang, M.: A constrained competitive swarm optimiser with an svm-based surrogate model for feature selection. IEEE Transactions on Evolutionary Computation (2022, DOI: 101109/TEVC20223197427)
14. Osaba, E., Del Ser, J., Martinez, A.D., Lobo, J.L., Nebro, A.J., Yang, X.S.: Momfcga: Multiobjective multifactorial cellular genetic algorithm for evolutionary multitasking. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–8. IEEE (2021)
15. Segev, N., Harel, M., Mannor, S., Crammer, K., El-Yaniv, R.: Learn on source, refine on target: a model transfer learning framework with random forests. IEEE Transactions on pattern analysis and machine intelligence **39**(9), 1811–1824 (2016)
16. Shi, Y., Lan, Z., Liu, W., Bi, W.: Extending semi-supervised learning methods for inductive transfer learning. In: 2009 Ninth IEEE international conference on data mining. pp. 483–492. IEEE (2009)
17. Wang, X.h., Zhang, Y., Sun, X.y., Wang, Y.l., Du, C.h.: Multi-objective feature selection based on artificial bee colony: An acceleration approach with variable sample size. Applied Soft Computing **88**, 106041 (2020)
18. While, L., Hingston, P., Barone, L., Huband, S.: A faster algorithm for calculating hypervolume. IEEE Transactions on evolutionary computation **10**(1), 29–38 (2006)
19. Xu, H., Xue, B., Zhang, M.: A duplication analysis-based evolutionary algorithm for biobjective feature selection. IEEE Transactions on Evolutionary Computation **25**(2), 205–218 (2020)
20. Xu, Q., Wang, N., Wang, L., Li, W., Sun, Q.: Multi-task optimization and multi-task evolutionary computation in the past five years: A brief review. Mathematics **9**(8), 864 (2021)
21. Zhang, F., Mei, Y., Nguyen, S., Zhang, M.: Multitask multiobjective genetic programming for automated scheduling heuristic learning in dynamic flexible job-shop scheduling. IEEE Transactions on Cybernetics (2022, DOI: 101109/TCYB20223196887)
22. Zhang, F., Mei, Y., Nguyen, S., Zhang, M., Tan, K.C.: Surrogate-assisted evolutionary multitask genetic programming for dynamic flexible job shop scheduling. IEEE Transactions on Evolutionary Computation (2021, DOI: 101109/TEVC20213065707)
23. Zhang, Q., Li, H.: Moea/d: A multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on evolutionary computation **11**(6), 712–731 (2007)
24. Zitzler, E., Laumanns, M., Thiele, L.: Spea2: Improving the strength pareto evolutionary algorithm. TIK-report **103** (2001)